Novel Methods in Computational Analysis and Design of Protein–Protein Interactions: Applications to Phosphoregulated Interactions

by

Brian Alan Joughin

Bachelor of Arts in Biophysics
Johns Hopkins University, 2000

Submitted to the Department of Biology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2007

Author .....
Department of Biology
September 21, 2006

Certified by .....
Bruce Tidor
Professor of Biological Engineering and Computer Science
Thesis Supervisor

Certified by .....
Michael B. Yaffe
Associate Professor of Biology and Biological Engineering
Thesis Supervisor

Accepted by .....
Stephen P. Bell
Chairman, Department Committee on Graduate Students

Novel Methods in Computational Analysis and Design of Protein–Protein Interactions: Applications to Phosphoregulated Interactions

by

Brian Alan Joughin

Submitted to the Department of Biology
on September 21, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biology

Abstract

This thesis presents a number of novel computational methods for the analysis and design of protein–protein complexes, and their application to the study of the interactions of phosphopeptides with phosphopeptide-binding domain interactions. A novel protein–protein interaction type, the action-at-a-distance interaction, is described in the complex of the TEM1 β-lactamase with the β-lactamase inhibitor protein (BLIP). New action-at-a-distance interactions were designed on the surface of BLIP and computed to enhance the affinity of that complex. A new method is described for the characterization and prediction of protein ligand-binding sites. This method was used to analyze the phosphoresidue-contacting sites of known phosphopeptide-binding domains, and to predict the sites of phosphoresidue-contact on some protein domains for which the correct site was not known. The design of a library of variant WW domains that is predicted to be enriched in domains that might have specificity for "pS/pT-Q" peptide ligands is detailed. General methods for designing libraries of degenerate oligonucleotides for expressing protein libraries as accurately as possible are given, and applied to the described WW domain variant library.

Thesis Supervisor: Bruce Tidor
Title: Professor of Biological Engineering and Computer Science

Thesis Supervisor: Michael B. Yaffe
Title: Associate Professor of Biology and Biological Engineering

Acknowledgments

I would like in particular to thank my wife, Clara, who has supported me throughout my graduate career, and who married me just in time to see me finish it. I also thank my parents and brothers, who have been very supportive.

I am deeply indebted to my advisors. I thank Bruce Tidor for providing a solid background in theory and computation, Michael Yaffe for experimental experience and for helping to keep my work motivated by interesting biology, and both for fostering significant interdisciplinary collaboration. I am also grateful to my other thesis committee members, Bob Sauer and Amy Keating, for helpful advice and suggestions.

The work in Chapter 2 of this thesis on action-at-a-distance interactions was done in collaboration with David Green, a former member of the Tidor group, and I am grateful to David for his insight into this and other work.

The idea of oligonucleotide design as a linear programming problem, and the formulations of the linear programs in Chapter 5, come from discussion with Mala Radhakrishnan, a student in the Tidor group.

I am additionally grateful to Shaun Lippow, Michael Altman, and other members of the Tidor group for numerous helpful discussions and advice.

I would like to thank members of the Yaffe group, and particularly Christian Reinhardt, Duaa Mohammad, Drew Lowery, Dan Lim, and Mary Stewart for helping me learn to do experimental biology. In particular, Christian Reinhart has been willing to think about my experiments as if they were his own, and his help has been invaluable to me. I would also like to thank Dane Wittrup, Daša Lipovšek, and Shaun Lippow from the Wittrup group for sharing reagents and advice related to yeast cell surface display.

Finally, I thank Mathew Corthell, Daniel Rinehart, and John Cataldo, for a great deal of barbecue- and game-related distraction.

Contents

1 Introduction **10**

2 Action-at-a-Distance Interactions Enhance Protein Binding Affinity 24

2.1 Introduction 25

2.2 Results and Discussion 26

2.3 Materials and Methods 34

2.4 Acknowledgments 36

3 A Computational Method for the Analysis and Prediction of Protein:Phosphopeptide-Binding Sites **37**

3.1 Introduction 38

3.2 Results 40

3.2.1 Phosphoresidue contact site properties 41

3.2.2 Predictive ability for known phosphoresidue contact sites . . . 44

3.2.3 Prediction of the phosphoresidue contact sites of Chk1 kinase and the BRCA1 BRCT-repeat domain 48

3.3 Discussion 51

3.4 Materials and Methods 53

3.4.1 Structures 53

3.4.2 Propensity calculation 54

3.4.3 Surface and contact calculation 55

3.4.4 Amino acid identity assignment 55

3.4.5 Mean surface curvature assignment 55

3.4.6 Solvated electrostatic potential assignment 55
3.5 Acknowledgments 56

4 Computational Design of a Library of WW Domain Variants Targeting an Altered Ligand Specificity 57
4.1 Introduction 58
4.2 Materials and Methods 60
4.2.1 Structure Preparation 60
4.2.2 Individual Protein Design 61
4.2.3 Protein Library Design 63
4.3 Results and Discussion 65
4.3.1 Individual Protein Design 65
4.3.2 Protein Library Design 70
4.4 Conclusion 72

5 Designing Optimally Small Degenerate DNA Libraries for Accurate Expression of Protein Libraries 74
5.1 Introduction 75
5.2 Methods 79
5.2.1 Protein library design 79
5.2.2 Solution of linear integer programs 80
5.2.3 Design of oligonucleotide libraries to represent protein libraries exactly 80
5.2.4 Design of oligonucleotide libraries with representation of undesired proteins allowed 82
5.2.5 Incorporation of cloning strategy-dependent information . . . 84
5.3 Results 86
5.4 Discussion 90

6 General Conclusions 93

A Testing the specificity of the Pin1 WW Domain **99**

A.1 Introduction 99

A.2 Materials and Methods 100

A.2.1 Cloning of the WW domain into pCT-CON2 in EBY100 yeast 100

A.2.2 Cell surface expression of the Pin1 WW domain 100

A.2.3 FACS analysis of surface-displayed WW domains 100

A.2.4 ELISA analysis of Pin1 WW domain specificity 101

A.3 Results and Discussion 102

A.3.1 FACS analysis of surface-displayed WW domains 102

A.3.2 ELISA analysis of Pin1 WW domain specificity 105

List of Figures

1.1 Structures of phosphopeptide-binding domains 16

2.1 Structure and residual potential of the TEM1-BLIP complex 28
2.2 Variation of experimental binding free energies with R 32

3.1 Structures of known phosphopeptide binding sites 39
3.2 Calculation of phosphoresidue contact propensities from global and phosphoresidue contact probability distributions 41
3.3 Calculation of joint propensity for phosphoresidue contact 45
3.4 Cross-validation of phosphoresidue contact site predictions on known phosphopeptide binding domains 47
3.5 Additional phosphoresidue contact site predictions 48
3.6 Predicted phosphoresidue contact sites on the surfaces of Chk1 and BRCA1 49

4.1 Library Design Method 64
4.2 Correlation of low- and high-accuracy energy evaluations 66
4.3 Characterization of designed WW variants by low- and high-accuracy energy 68

5.1 Complications in Combinatorial Oligonucleotide Library Design . . . 76
5.2 Oligonucleotide Library Design 78
5.3 Linear Programming 79
5.4 Cloning Strategy 84
5.5 Size and content of designed oligonucleotide libraries 88

A.1 The Pin1 WW domain is cell surface expressed 103
A.2 The surface expressed Pin1 WW domain does not bind "pT-P" peptide library specifically 104
A.3 A GST-Pin1 WW domain fusion is specific for "pS/T-P" peptides and peptide libraries 106

List of Tables

1.1 Phosphopeptide-binding domains and their specificities 15

2.1 Energetic details of mutations to BLIP 30

3.1 Structures used to calculate propensity data 54

4.1 Characterization of wild-type Pin1 compared to designed WW variants 65

4.2 Designed "pS/pT-Q"-binding library 72

5.1 The fifteen degenerate nucleotide mixtures 75

5.2 Size and content of oligonucleotide libraries 87

5.3 A designed oligonucleotide library 91

Chapter 1

Introduction

Phosphoproteins are of tremendous importance in eukaryotic cellular signaling and regulation. Protein kinases are the 3rd most common family of protein in the human genome, with 575 members [1]. It has been estimated that 30% of human proteins are substrates for these kinases [2], with resulting effects in nearly every aspect of biology.

Protein kinases generate phosphopeptides and phosphoproteins

Protein kinases catalyze the transfer of a phosphate moiety from ATP to a protein. Although other residues can be phosphorylated, the vast majority of stable protein phosphorylations in eukaryotes occur on the hydroxyl moiety of serine, threonine, and tyrosine amino acid side chains. An analysis of several eukaryotic genomes suggests that 75% of eukaryotic kinases phosphorylate serine and threonine residues, while the remaining 25% are tyrosine-specific [3]. Phosphorylation of proteins by protein kinases is regulated at several levels. First, the kinase and potential ligand must be co-localized in the cell. Second, the kinase must be in an active form. The catalytic domains of most protein kinases share a similar active state conformation, but are often activated by a conformational change from a less similar inactive state, driven by a variety of mechanisms [reviewed in 4]. These mechanisms include phosphorylation of the kinase itself on a flexible "activation" loop [5–7], inter- or intramolecular allostery [8–10], and release of an intramolecular substrate-competitive inhibitor sequence [11–15]. It has been noted [4] that the diversity among kinase domain inactive state conformations lends itself to the design of kinase inhibitors that function specifically

by stabilization of the kinase-specific inactive states, as the anticancer drug Gleevec does in the case of the Abl kinase [16, 17], rather than less specifically by obstructing the relatively well-conserved nucleotide-binding site. Finally, kinase activity is also regulated by recognition of substrate amino acid sequence surrounding the phosphate receptor side-chain [18].

In general, the experimental demonstration that a particular protein kinase phosphorylates a particular protein substrate is onerous. There are three generally-accepted requirements [19]. First, the kinase must be capable of phosphorylating the putative substrate *in vitro* with physiologically relevant K_m and v_{max}. Second, the substrate must be phosphorylated *in vivo* at the same amino acid side chain in response to a signal that activates the protein kinase. Finally, since there are kinases that share both activating signal and substrate specificity, the substrate should be shown not to be phosphorylated when the kinase of interest is specifically inactivated, either chemically, genetically, or at the transcriptional or translational level. Because of the difficulty of such a characterization, experimental tools have been developed for generating hypotheses connecting phosphoproteins and kinases.

Identifying the products of a kinase of interest

There are two families of techniques for identifying potential downstream products of an individual kinase. The first is direct: individual proteins are identified that are phosphorylated, generally *in vitro*, as a response to the activity of a kinase of interest. The laboratory of Kevan Shokat has made notable advances in this area, designing variant kinases that accept as a phosphate donor a radiolabeled analog of ATP that does not interact with wild-type kinases [20–23]. The analog is added to cell lysates, and radiolabeled proteins, putatively phosphorylated by the variant kinase, are subsequently identified. Other "direct" approaches include affinity reagents such as phosphomotif antibodies that can be used in some cases to purify phosphorylated peptides and proteins [24–26] for identification. A number of other techniques exist for querying the phosphorylation state of the proteome more globally that can be used to generate hypothetical kinase-substrate relationships [27] for further testing,

including 2-D gel electrophoresis, and mass spectroscopy.

The second set of techniques is more indirect. A motif-based characterization of the substrate sequence specificity of the kinase is generated, and that characterization is used to search a protein sequence database to generate hypotheses. Phosphorylation motifs have been determined by peptide library screening [28, 29], in which a pool of degenerate peptides containing a fixed, or "oriented", serine, threonine, or tyrosine is exposed to a kinase, and the phosphorylated subset of the pool is purified by chromatography and batch sequenced to provide a consensus sequence. In a modification to this technique, libraries of degenerate phosphopeptides have been spotted on membranes, and exposed to a kinase of interest and radiolabeled ATP [30]. The spotting allows for parallelization; libraries can be searched with every amino acid fixed at every positon relative to the fixed phosphate acceptor residue in a high throughput manner. In an alternative approach, peptide libraries have been immobilized on beads, exposed to a kinase, and then sorted by FACS after incubation with a phosphospecific fluorophore [31]. Libraries containing fusions of peptides to the mRNAs that encode them have also been screened by immunoprecipitation with a phosphospecific antibody, followed by analysis on a cDNA microarray [32].

Once the phosphorylation motif for a particular kinase is determined, database techniques can be applied to scan the set of known protein sequences for matches, which can all be considered for further testing as putative kinase targets. The method of Yaffe *et al.* [33, 34], implemented in the Scansite program, scores potential phosphorylation sites by matching protein sequences against a matrix of kinase motif selectivity values generated by library screening, and provides a measure of statistical relevance by percentile-ranking the site in question among all potential sites in known protein sequences. It is possible to discover a great number of leads for experimental validation quickly in this manner, though the identification of motif sequences by a database search provides no direct experimental evidence for the phosphorylation of target proteins.

Identifying the kinase generating a site of interest

There is, to date, no method of identifying the kinase responsible for a phosphorylation of interest from the entire kinome without some prior hypothesis. Such a hypothesis can be generated based on knowledge of what kinases are active under the circumstances and in the location in which the site of interest is phosphorylated, on knowledge of what kinases have motif specificity compatible with the phosphorylated sequence [33, 34], or on knowledge of what antagonists inhibit the phosphorylation of interest.

In vivo function of protein phosphorylation

Induction of conformational change

The steric bulk and strong anionic charge of the phosphate ion can prompt a conformational change in a peptide upon phosphorylation. The first known example of conformational change as the effect mechanism of phosphorylation was in the protein glycogen phosphorylase [35]. Phosphorylation of glycogen phosphorylase at residue serine 14 prompts a local conformational rearrangement that leads to large-scale allosteric shift of both the monomeric and multimeric protein structure, resulting in the activation of the enzyme. The activities of some protein kinases, including IRK [5, 36] and the MAP kinases ERK2 [6] and p38 [37] are also regulated by a phosphorylation-dependent conformational change in the kinase activation loop, again resulting in a conformational change that results in enzyme activation.

Generation of a docking site for a phosphopeptide-binding domain

Rather than a conformational change, it is possible for protein phosphorylation to affect protein function by the creation of a binding site for a phosphopeptide-binding domain [see 38–42, for reviews]. Generally, phosphopeptide-binding domains have

specificity for a phosphorylated amino acid, as well as at least partial specificity for some amino acids in a short sequence motif surrounding the phosphorylated residue.

In 1990, it was recognized in the research groups of Tony Pawson and Hidesaburo Hanafusa that a noncatalytic domain conserved among several tyrosine kinases, including Src, Abl, and Fps, was responsible for binding to receptor tyrosine kinases, in a manner directed by the autophosphorylation of those kinases [43–47]. This domain, named the "*Src* homology region 2", or SH2 domain, was the first discovery of a growing number of proteins and protein domains that bind phosphorylated proteins and peptides through a short sequence motif in the surrounding amino acids. An energetics-based analysis of the SH2 domain of the proto-oncogenic tyrosine kinase Src, showed that half of free energy of binding of a high-affinity peptide came from binding the phosphotyrosine itself, and that the rest of the peptide conferred the other half [48]. The next domain to be characterized, the PTB, or "phosphotyrosine-binding" domain [49, 50], was also shown to bind phosphotyrosine-containing peptides specifically, though with differences in the surrounding motif. It was thought for some time later, therefore, that while tyrosine phosphorylation could result in either a conformational change, or the generation of a phosphotyrosine-binding protein docking site, phosphorylated serines and threonines acted solely through conformational change.

This changed with the identification of the ubiquitous 14-3-3 proteins as a phosphoserine binding domain in 1996 [51, 52]. There are 7 genetically distinct mammalian 14-3-3 isotypes, with a great deal of sequence similarity [53]. More than 50 substrates for members of the family have been identified to date, with wide-ranging effects. Since the discovery of the capacity of 14-3-3 to bind phosphopeptides, a number of other phosphoserine- and phosphothreoinine-binding domains, including the WW domain [54] of the phosphodirected proline isomerase Pin1 [55], the FHA domain [56], the Polo-box domain [57], and the BRCT domain [58] have been identified. It seems certain that more domains remain to be found. In particular, the existence of the WW domain, a family of peptide binding domain within which only a subset are phosphospecific, seems to indicate that perhaps even modular domains that are al-

Table 1.1: **Phosphopeptide-binding domains and their specificities.**

Domain	Example	Specificity	Reference
SH2	Src	pY-E-E-I	[59]
PTB	SHC	N-P-x-pY	[60]
14-3-3	14-3-3 ζ	R-(S/Ar)-X-(pS/pT)-X-P R-X-(Ar/S)-X-(pS/pT)-X-P	[52]
Group IV WW	Pin1	(pS/pT)-P	[54]
FHA	Rad53 FHA1	pT-X-X-D	[61]
WD40	β-TrCP	D-pS-G-X-X-pS	[62]
MH2	Smad2	S-pS-M-pS-COOH	[63]
Polo-box	Plk1	S-(pS/pT)-P	[64]
BRCT	BRCA1	(pS/pT)-X-X-(F/Y)	[58]

ready known, such as the SH3 and PDZ domains, will eventually be shown to have some phosphospecific members.

The phosphopeptide-binding domains that have been discovered have disparate specificities, and differing modes of phosphate coordination. A list of the known phosphopeptide-binding domains, and their canonical specificities, is given as Table 1.1. Several phosphopeptide-binding domain structures are shown in Figure 1.1, showing the lack of similarity in the structure of the domains, both in the global fold and in the mechanism by which phosphopeptides are bound.

A number of roles have been discovered for phosphopeptide-binding domains in a range of important biological processes from apoptosis to cell-cycle control and differentiation. Some of the isoforms of the phosphopeptide-binding protein 14-3-3 exert a tumor-suppressing function in the DNA damage pathway by isolating the pro-mitotic phosphatase Cdc25 in the cytosol when Cdc25 has been phosphorylated by the DNA damage kinase Chk1 [68–72]. The *S. cerevisiae* protein Ess1, the budding yeast homologue of Pin1, contains a WW domain that exerts regulatory effects in transcription by phosphodependently binding the C-terminal domain of RNA polymerase II, which is then isomerized by a second, catalytic, domain on Ess1 [73]. The Polo-box domain of Plk family kinases, which regulate aspects of mitosis and cytokinesis [74], is required for protein localization, and likely for substrate targeting as

A. 14-3-3 ζ **B.** Pin1 WW domain

C. Rad53 FHA1 domain **D.** p56^{Lck} SH2 domain

Figure 1.1: **Structures of phosphopeptide-binding domains.** Backbone structures of (A) a dimer of 14-3-3 ζ [65], (B) the Pin1 WW domain [66], (C) the FHA1 domain of Rad53 [61], and (D) the p56^{Lck} SH2 domain [67] in complex with cognate phosphopeptides. In all cases, the domain backbone is colored blue, and the peptide backbone is colored red. Note the dissimilarity in secondary and tertiary structure.

well by binding to peptides containing the motif "S-(pS/pT)-P" [57, 75]. A defect in the wild-type ability of the BRCA1 BRCT domain to bind phosphopeptides [58, 76] predisposes women to breast and ovarian cancer [77, 78]. BRCA1 mutations linked to inherited cancer-related phenotypes are enriched in the BRCT domains of the protein. In conjunction with kinases and phosphatases with which they have fully or partially overlapping ligand specificities, phosphopeptide-binding domains are capable of exerting combinatorial control over cell signaling by temporally and spatially controlling the assembly and disassembly of signaling complexes.

Identification of the substrates of a phosphopeptide-binding domain

The process of identifying the substrates of a phosphopeptide-binding domain is quite similar to the identification of kinase substrates. Targets can be identified directly, by affinity purification using as bait the domain of interest [79]. Oriented peptide library screening has been used extremely successfully as a more indirect technique to first identify in batch all peptides from a random library that bind to a domain of interest [52, 80], and then use the consensus profile to predict *in vivo* ligands [33, 34].

Identifying phosphopeptide-binding domains specific for a site of interest

The two most recently described phosphopeptide binding domains, the Polo-box domain [57] and the BRCT domain [58], were identified using a proteomic screen for proteins that bind to a target peptide library phosphospecifically. Proteins were translated *in vitro* in pools, and pools were tested for binding to phosphorylated and nonphosphorylated peptide libraries. Members of pools that showed specificity were searched for specific clones responsible for the activity. The Plk-1 Polo-box was identified in a search for proteins specific to the motif "pS/pT-P" generated by mitotic kinases. The WW domain of Pin1, known to share that specificity, was also re-identified. The BRCT domains of BRCA1 and PTIP were identified in an equivalent screen for domains that bound the motif "pS/pT-Q", generated by the DNA damage kinases. Interestingly, though in the optimal motif for the BRCT domain, a phenylalanine at the pS/pT+3 position seems to be much more important than a

pS/pT+1 glutamine. In both cases, the experimental design was driven by a desire to understand the downstream regulation prompted by the activity of a particular kinase family; however, neither domain had identical specificity to that of the targeted kinase. Kinases and phosphatases, by having incompletely-overlapping specificity, are able to exert a combinatorial control in signaling.

Disruption of sequence-specific effects

In some cases, the effect of phosphorylation is to disrupt some other sequence-mediated function. It has been shown, for example, that phosphorylation of 14-3-3 ζ by MAPKAP kinase 2 impairs the ability of 14-3-3 ζ to dimerize [81]. Phosphorylation of Cdc25C at serine 214 prevents phosphorylation at serine 216, preventing the binding and cytosolic sequestration of Cdc25C by 14-3-3 [82]. Phosphorylation of the Forkhead transcription factor AFX at serine 193 in the middle of a nuclear localization sequence motif results in an increase of cytosolic AFX, and a decrease in AFX transcriptional activity [83]. Because phosphorylation results in a significant change in the steric and electrostatic properties of any local peptide sequence, it is quite easy to imagine the direct disruption of nearly any nonphosphorylated sequence motif-mediated effect by phosphorylation. Function can also be altered indirectly, through binding a phosphopeptide binding domain. 14-3-3 binds to the transcription factor FKHRL1, and in so doing exposes on FKHRL1 a nuclear export sequence, prompting it to leave the nucleus [84].

Protein phosphatases remove protein phosphorylation

Protein phosphorylations are removed by the action of protein phosphatases, providing the aspect of reversability to phosphoregulation. Protein phosphatases are fewer in number than protein kinases [3], and demonstrate a much broader specificity *in vitro*. *In vivo*, regulation and targeting of phosphatases is handled, in at least some

cases (including the protein phosphatase 1 family of serine/threonine phosphatase), by the complexing of phosphatase with other proteins to form a number of separate holoenzymes [2, 85, 86]. Many protein tyrosine phosphatases, on the other hand, are found in tandem with other protein domains that seem to have the capacity for generating target specificity through localization or direct binding through SH2 and PTB domains [87]. Interestingly, this modular multidomain structure is similar to that found in protein tyrosine kinases. Unlike serine/threonine phosphatases, there are similar numbers of tyrosine kinases and phosphatases [24], indicating that perhaps in the case of phosphotyrosine regulation, there is more mechanism shared between kinases and phosphatases.

The present work

In this thesis, we have developed general methods for the analysis and design of protein–protein interactions. We have applied these methods in particular to the interaction of phosphopeptide-binding domains with their cognate phosphopeptides. The work described here brings together research in a number of different fields including computational biophysics, biochemistry, molecular biology, computer science, mathematics, and biotechnology, as needed to make headway against the problems encountered. There is experimental and theoretical research described here; it is our feeling that each is more powerful when informed by the other.

Protein–protein interactions are governed by a balance of the energetics between the proteins involved, in the bound state, and between the individual proteins and their environment in the unbound state. Because water–water interactions are, by and large, more favorable than the water–protein interactions with nonpolar atoms at the surface of a protein molecule, proteins are generally driven to reduce their surface area [88]. For a single molecule, this is the process of protein folding; multiple folded proteins can likewise reduce their total surface area by forming a multimolecular complex. The details of when and how these processes occur depend minutely on the energetic details of solvent–solute and solute–solute interactions.

A variety of novel methods are developed and employed in this work to solve problems of computational protein–protein complex analysis and design with a focus on phosphodependent interactions. With a single exception, the problems solved were motivated by a desire to understand or design the interactions phosphopeptide-binding domains with their cognate phosphopeptides in a phosphorylation-dependent manner.

In Chapter 2 of this thesis, we have proposed a novel description of a type of protein–protein interaction that we term the action-at-a-distance interaction. A study of the wild-type complex of the TEM1 β-lactamase with the β-lactamase inhibitor protein (BLIP), along with some previously-characterized mutants [89] revealed the existence of a region of the surface of BLIP, but away from the TEM1–BLIP interface, on which negatively-charged amino acids may interact particularly favorably with TEM1. These charged residues lie outside the direct binding interface, paying a negligible desolvation penalty, but are capable nonetheless of interacting strongly, for 1 kcal/mol or more of favorable binding free energy.

While protein design typically entails in-depth consideration of the atomic-scale details of molecular interactions, the action-at-a-distance interaction suggests a more subtle, but possibly more forgiving means of introducing mutations outside of protein–protein binding interfaces in an appropriate manner to resolve unsatisfied residual electrostatic potential within the interface. We have identified several novel surface mutations to BLIP that we expect to improve affinity to TEM1. These mutations do not make any intimate contact with TEM1, and some are further than 7 Å from the closest TEM1 atom.

In Chapter 3, we attempted to gain an understanding of the nature of phosphopeptide binding via the structural analysis of a set of known phosphopeptide-binding domains. Though the biochemical bases for recognition of the phosphorylated ligand side chain vary greatly among the domains, we theorized that some chemical and physical characteristics must nonetheless be shared. A machine learning method was used to obtain an understanding of the chemical and physical properties important in the recognition of phosphorylated amino acids by phosphopeptide-binding domains.

A framework was developed in which chemical and physical characteristics are evaluated at the vertices of finely-discretized meshes describing the surfaces of the known phosphopeptide-binding domains. The capacity of these characteristics to differentiate the phosphoresidue contact surfaces of the domains from the surfaces in general was evaluated. Jack-knife validation indicated that a mathematical model based on the calculation of enrichment propensities of physical and chemical characteristics was capable of identifying the phosphoresidue contact site of a phosphopeptide-binding domain not used to train the model.

As a test of its utility, we used this model to identify putative phosphoresidue contact sites on the surfaces of two phosphopeptide-binding domains for which the site was not known at the time of the work, the Chk1 kinase domain and the BRCA1 BRCT domain. One of two predictions on the Chk1 kinase domain was in good agreement with published biochemical data [90], while several groups have published crystal structures [78, 91, 92] in agreement with one of the two predictions made on the BRCT domain surface. The method described is not limited to use in the prediction of phosphopeptide-binding sites, but can be applied to any protein–ligand interaction type. In particular, the evaluation of properties on a very fine mesh can be an aid in situations where few crystal structures of complexes are available for use in training, although there are obvious dangers here in overfitting data. Moreover, we have developed a framework within which *any* chemical or physical property, either discrete or continuous in nature, which can be evaluated at a point on the surface of a molecule can be evaluated for its contribution to the likelihood of binding a particular ligand.

The development of variant phosphopeptide-binding domains with putative novel specificity is described in Chapter 4. The design target was a WW domain capable of binding to protein products phosphorylated on "S/T-Q" sequence motifs by the kinases ATM and ATR [reviewed in 93, 94], which phosphorylate ligands as an early response to DNA damage by IR and UV irradiation. Such a domain would be a valuable laboratory reagent for identifying unknown, and potentially therapeutically relevant, new proteins in the DNA damage signaling pathway. Traditional single-

protein design of a mutant Pin1 WW domain capable of forming a stable complex with peptides containing the phosphorylated sequence motif "pS/pT-Q", rather than the wild-type Pin1 target "pS/pT-P", gave results that looked unlikely to function as desired *in vitro*.

A WW domain variant library was designed based on the hypothesis that any domain that binds tightly and specifically to the glutamine residue of the "pS/pT-Q" motif must do so through hydrogen bonding. The guiding philosophy in the development of this library was that the library itself should be liberal in including all protein sequences that might have the desired specificity. In practice, that means that we have decided to rely on experimental screening to identify those sequences that can make flexible adjustments in backbone structure in order to form the desired complex. We have used computation only to determine for what protein sequences such a backbone relaxation might be plausible. The designed library contains every variant WW domain sequence that might be capable of making three or more hydrogen bonds to the peptidyl glutamine ligand. The ongoing development of an experimental system for screening this library, based on the yeast cell surface display system developed by Boder and Wittrup [95] is described in Appendix A.

Finally, in Chapter 5, a suite of tools was developed for the design of oligonucleotide libraries based on the well-studied mathematical optimization method of linear integer programming [96]. A method is described for finding the smallest set of degenerate oligonucleotides that encode an arbitrary list of protein sequences exactly, with no extras. An extension is described that allows for the generation of smaller oligonucleotide libraries by allowing the inclusion of a user-selected number of undesired sequences. A related extension can be used to generate a list of degenerate oligonucleotides of a given size that encodes all desired proteins, and as few undesired proteins as possible. We also give a sample procedure for using known information about the method by which an oligonucleotide library will be expressed to inform the design of that library.

These methods were applied to the design of oligonucleotide libraries that express the protein library of WW domain variants described in Chapter 5, where we found

that surprisingly small sets of degenerate nucleic acids could encode all members of our protein library, with only a small number of undesired additional protein sequences. Though the library contained over 100,000 protein sequences, a set of 69 double-stranded oligonucleotides could encode it exactly with no extra full-length proteins. A smaller library of 23 double-stranded degenerate oligonucleotides can be designed that encodes all desired proteins, and a roughly equal number of undesired protein sequences. The most powerful improvements in oligonucleotide library size come from accounting for the experimental method by which the library will be expressed. A library of 64 *single-stranded* oligonucleotides can exactly encode the desired protein library, while a smaller library of only 50 single-stranded oligonucleotides encodes the desired protein library, along with approximately 10% more undesired protein sequences. This can be compared with the traditional combinatorial oligonucleotide library, which is a double-stranded oligonucleotide that encodes 3.2×10^6 proteins, only 3% of which are specified by the target protein library.

In this thesis, we have attempted to gain a better understanding of protein–protein interactions in general, and specifically of the binding of phosphopeptides to phosphopeptide-binding domains. We have identified a novel interaction type, the action-at-a-distance interaction, and designed several such interaction that we propose will enhance the affinity of the TEM1/BLIP protein complex. We have performed an analysis of the important determinants of phosphopeptide binding, and used the resultant model to predict the phosphopeptide binding site on at least one phosphopeptide-binding domain, the BRCA1 BRCT domain. We have designed a library of variant phosphopeptide-binding domains that may be enriched for domains with a novel specificity for "pS/pT-Q"-containing peptide ligands, and are working toward experimental screening of this library. We have developed tools for designing degenerate oligonucleotide libraries that can express this and other protein libraries with higher fidelity than combinatorial libraries, the current standard. We expect the methods developed to be of broad utility beyond the field of phosphopeptide-binding domain interactions, to the analysis and design of protein–protein interactions in general.

Chapter 2

Action-at-a-Distance Interactions Enhance Protein Binding Affinity [1]

Abstract

The identification of protein mutations that enhance binding affinity may be achieved by computational or experimental means, or by a combination of the two. Sources of affinity enhancement may include improvements to the net balance of binding interactions of residues forming intermolecular *contacts* at the binding interface, such as packing and hydrogen-bonding interactions. Here we identify *non-contacting* residues that make substantial contributions to binding affinity and that also provide opportunities for mutations that increase binding affinity of the TEM1 β-lactamase (TEM) to the β-lactamase inhibitor protein (BLIP). A region of BLIP not on the direct TEM1-binding surface was identified for which changes in net charge result in particularly large increases in computed binding affinity. Some mutations to this region have previously been characterized [89], and our results are in good correspondence with this results of that study. In addition, we propose novel mutations to BLIP that were computed to improve binding significantly without contacting TEM1 directly. This class of non-contacting electrostatic interactions could have general utility in the design and tuning of binding interactions.

[1] This chapter has previously been published as:

Brian A. Joughin, David F. Green, and Bruce Tidor. Action-at-a-distance interactions enhance protein binding affinity. *Protein Science*, 14:1363–1369, 2005.

2.1 Introduction

The field of protein design has made substantial advances over the last twenty years, based largely on phrasing the appropriate inverse problem and developing methods capable of addressing inverse design [97, 98]. Much current protein design work involves the construction of stabilizing protein side-chain arrangements by methods such as dead-end elimination [99–107], self-consistent mean-field theory [108–111], simulated annealing [112–116], genetic algorithms [117–121], and combinatorial search [117–119]. That is, successful design has been achieved by consideration of detailed atomic interactions and their effect on packing geometry and energetics [122–125]. The design of protein binding interfaces may be achieved by a similar overall approach, although the additional requirement to treat solvation and electrostatic interactions adds a further layer of complexity [126].

An alternative strategy that does not demand the same detailed packing of side chains into an exquisite three-dimensional jigsaw puzzle may be desirable in many cases. One such method involves the enhancement of affinity through relatively long-range electrostatic effects by the mutation of surface residues located somewhat outside of the binding interface. When *surface* mutations are not located directly at the binding interface, a detailed consideration of packing may be unnecessary. Moreover, when the effects of mutations act over a relatively long range, such as through electrostatic interactions, design attributes should be more tolerant of local imperfections in structural models. Less apparent, however, is how effective these types of mutations can be (since much of the interaction may be screened by solvent), and how particularly favorable mutations of this class can be identified. An important design consideration is the counterplay of favorable intermolecular electrostatic interactions made between the partners in the bound state and unfavorable desolvation costs incurred by each partner due to binding; this balance leads to counterintuitive behavior for the energetics of electrostatic interactions in biological systems [e.g., 127–129]. The lessons learned from detailed analyses of short-range electrostatic interactions, such as salt-bridges and hydrogen-bond networks, may or may not prove to be extendible

in a straight-forward manner to longer-range electrostatic interactions of this nature (termed here "action-at-a-distance" interactions).

2.2 Results and Discussion

We have begun to address these issues by analyzing the affinity of the β-lactamase inhibitor protein (BLIP) for binding the TEM1 β-lactamase, with a focus on electrostatic interactions. Using methods based on a continuum solvation model, we computed the electrostatic contributions to the energetics of TEM1 binding for wild-type BLIP and for a set of BLIP mutants whose changes were focused at surface positions. The degree of electrostatic complementarity between binding partners correlates well with the experimentally determined binding affinities, which suggests that these complementarity tools may be particularly useful both in understanding and in designing surface mutations. To complete the binding analysis, both van der Waals and hydrophobic contributions to the binding energetics were also calculated. Preliminary analysis indicated that change in side-chain entropy was not a significant component of binding energy for the residues examined in this study, and is not considered here.

Our laboratory has previously described a measure of electrostatic complementarity between two binding partners [130, 131]. Termed the residual potential and computed from continuum electrostatic calculations, this measure can be expressed numerically as a statistical quantity or viewed graphically overlaid on the structure, which highlights regions of particularly high or low electrostatic complementarity. The consideration of electrostatics in binding involves balancing favorable interactions made between the members of the complex in the bound state with the loss of favorable interactions that each component makes with solvent upon binding. For perfect complementarity, this balance is met such that the interaction potential of the receptor is opposite in sign and equal in magnitude to the ligand desolvation potential. Thus, we may derive a measure, termed the residual potential, that describes the balance:

$$\phi^{\text{resid}} = \phi_{\text{R}}^{\text{inter}} + \phi_{\text{L}}^{\text{desolv}} \tag{2.1}$$

The residual potential is near zero in regions of high complementarity and is larger in magnitude in regions of poorer complementarity. It is important to note that the definition of the residual potential is fundamentally asymmetric, describing the complementarity of the ligand for binding the receptor. A complex for which the ligand is perfectly complementary to its receptor may not be as complementary when the roles of its components are reversed; the receptor may not be perfectly complementary to the ligand [131]. Also, the definition here applies to binding with no conformational change. This is a reasonable approximation of TEM–BLIP binding, as the RMSD values for main-chain atoms of TEM and BLIP upon complex formation from the apo states [132, 133] are 0.35 and 0.70 Å, respectively [134]. A numerical statistic for the complementarity of a ligand for its receptor can be obtained from the correlation of the interaction and desolvation potentials, $\phi_{\mathrm{R}}^{\mathrm{inter}}$ and $\phi_{\mathrm{L}}^{\mathrm{desolv}}$,

$$R = \frac{\sum[(\phi_{\mathrm{R}}^{\mathrm{inter}} - <\phi_{\mathrm{R}}^{\mathrm{inter}}>)\cdot(\phi_{\mathrm{L}}^{\mathrm{desolv}} - <\phi_{\mathrm{L}}^{\mathrm{desolv}}>)]}{[\sum(\phi_{\mathrm{R}}^{\mathrm{inter}} - <\phi_{\mathrm{R}}^{\mathrm{inter}}>)^2 \cdot \sum(\phi_{\mathrm{L}}^{\mathrm{desolv}} - <\phi_{\mathrm{L}}^{\mathrm{desolv}}>)^2]^{1/2}} \tag{2.2}$$

where the summations run over the points of interest (typically sampling the molecular surface of the ligand) and quantities in angle brackets represent averages over the points. For perfect complementarity the correlation coefficient is -1. Negative values smaller in magnitude indicate imperfect complementarity, while positive values indicate anticomplementarity.

Wild-type BLIP binds to TEM1 with a K_d of 1.25 nM [89], burying 2560 Å^2 of solvent exposed surface and forming eleven hydrogen bonds and four salt-bridges across the binding interface, making it a fairly typical enzyme–inhibitor complex [137]. The residual potential for TEM1 binding on the surface of BLIP was computed and is displayed in Figure 2.1C, along with an overview of the structure in Figures 2.1A and B. The desolvation potential of BLIP is quite complementary to the interaction potential of TEM1 projected onto the BLIP surface; most regions of positive desolvation potential are well matched by regions of negative interaction potential, and vice versa. However, examination of the residual potential makes it clear that BLIP is not perfectly complementary to TEM1. Specifically, the net residual potential is

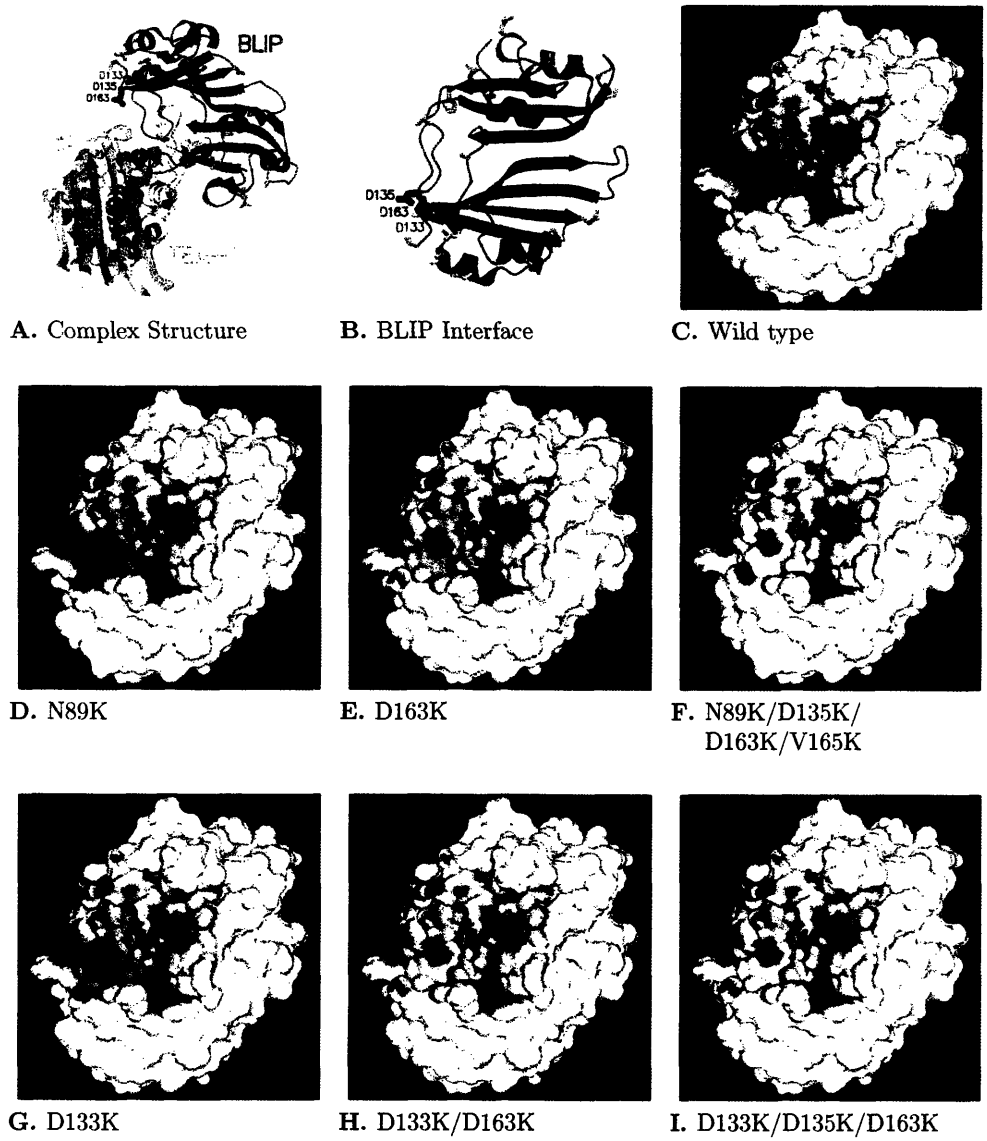


Figure 2.1: **Structure and residual potential of the TEM1-BLIP complex.** (A) Structure of the complex between BLIP and TEM1. Selected mutated side chains are included. Locations of high-activity mutations are labeled individually. Olive-colored residues indicate sites of low-activity mutations. Figure made with MOLSCRIPT [135] and RASTER3D [136]. (B) Structure of BLIP at the TEM1-BLIP interface, with residues shown as described in (A). (C)-(I) Residual potentials for TEM1-BLIP binding on the surface of BLIP variants: (C) wild-type; (D) N89K; (E) D163K; (F) N89K, D135K, D163K, V165K; (G) D133K; (H) D133K, D163K; (I) D133K, D135K, D163K. Residual potentials are colored on linear scales from 0 to -20 kT/e in red and 0 to 20 kT/e in blue.

negative over a large area of the binding surface. This can be viewed as either an excess negative interaction potential from TEM1 or as an insufficiently positive desolvation potential from BLIP. Thus, this suggests that the binding affinity of BLIP for TEM1 may be improved by appropriate mutations that increase the relative positive charge on the inhibitor (mutations of acidic residues to neutral or basic residues, and mutations of neutral residues to basic residues) or by mutations that decrease the relative negative charge on the enzyme. However, the asymmetry of the residual potential suggests that such mutations should be targeted to particular regions of the periphery of the interface.

In order to address the question of asymmetry in the effectiveness of mutations, a set of surface residues near, but not at, the the binding interface (the "periphery" of the interface) were chosen. The potential for improving the electrostatic complementarity of BLIP for TEM-1 by mutation of each of these residues to lysine was then investigated. Mutant structures were built, and their relative electrostatic complementarity and binding affinities were estimated computationally (see Table 2.1). The results show two classes of single mutants, "low activity" and "high activity". The low-activity mutations produced little change in $\Delta G_{\text{bind}}^{\text{comput}}$ and in R (the residual potential measure of electrostatic complementarity), with estimated enhancements in binding affinity of 1 kcal/mol or less relative to wild-type. When the residual potential was examined on a high-resolution computer graphics system, no change could be seen visually (see Figure 2.1D). By contrast, high-activity mutations resulted in significant changes in computed electrostatic complementarity and binding affinity. Three of the four BLIP amino-acid positions for which high-activity mutations were found, Asp 133, Asp 135 and Gln 161, are a sufficient distance from the TEM1 binding site that no solvent-accessible surface area is buried for these positions upon binding, and negligible changes in van der Waals binding free energy result (less than 0.1 kcal/mol relative to wild-type upon mutation to lysine). Nevertheless, computations predict that these mutations improve binding the free energy of the TEM/BLIP complex by 1.9, 1.4, and 1.2 kcal/mol, respectively. The other, Asp 163, does not contact TEM1 in the wild-type complex structure, but does in the computed model

Table 2.1: **Energetic details of mutations to BLIP.**

BLIP Mutations	R	$\Delta\Delta G_{es}$	$\Delta\Delta G_{vdw}$	$\Delta\Delta G_{SASA}$	$\Delta\Delta G_{calc}$	$\Delta\Delta G^{a}_{exp}$
Wild-type	-0.63	0.00	0.00	0.00	0.00	0.00
V3K	-0.64	-0.09	-0.05	0.00	-0.14	N/D
T5K	-0.64	-0.32	-0.02	0.00	-0.35	N/D
E28K	-0.64	-0.02	-0.01	0.00	-0.04	N/D
T32K	-0.64	-0.64	0.05	0.00	-0.59	0.19
H45K	-0.65	-0.41	0.01	0.00	-0.40	N/D
S60K	-0.63	-0.12	-0.01	0.00	-0.12	N/D
A61K	-0.63	-0.21	-0.08	0.00	-0.29	N/D
A77K	-0.64	-0.23	-0.02	0.00	-0.25	N/D
L85K	-0.64	-0.35	-0.01	0.00	-0.35	N/D
N89K	-0.63	-0.14	0.00	0.00	-0.14	-0.47
V93K	-0.64	-0.27	-0.03	0.00	-0.29	-0.49
V134K	-0.64	-0.41	-0.02	0.00	-0.42	N/D
T140K	-0.63	0.37	-0.54	-0.15	-0.32	-0.02
D153K	-0.64	-0.28	-0.01	0.00	-0.29	N/D
Q157K	-0.64	-0.28	0.00	0.00	-0.28	N/D
Q161K	-0.65	-1.15	-0.01	0.00	-1.16	N/D
V165K	-0.64	-0.64	-0.03	0.00	-0.67	N/D
D133K	-0.66	-1.85	-0.04	0.00	-1.88	N/D
D135K	-0.65	-1.32	-0.06	0.00	-1.38	N/D
D163A	-0.69	-5.00	0.93	0.17	-3.90	-1.34
D163K	-0.71	-4.46	-3.22	-0.55	-8.23	-1.99
T140K/Q157K	-0.64	-0.46	-0.54	-0.15	-1.14	-0.41
N89K/D163K/V165K	-0.76	-5.26	-3.12	-0.96	-9.34	-2.40
V134K/D135K/D163K	-0.73	-5.19	-3.29	-0.55	-9.03	-3.06
N89K/D135K/ D163K/V165K	-0.77	-5.64	-3.18	-0.96	-9.78	-3.36
D133K/D163K	-0.74	-5.63	-3.21	-0.554	-9.45	N/D
D133K/D135K/D163K	-0.75	-6.15	-3.22	-0.554	-9.87	N/D

[a]Experimental results from [89]

mutation to lysine. To determine whether the contact involving Lys 163 is important for its ability to stabilize the TEM/BLIP complex, we also studied Lys 163 in an extended conformation that is not significantly buried upon TEM1 binding. In both conformations we find calculated binding improvements of over 5 kcal/mol, relative to wild-type, although the minimized conformation is significantly more favorable than the extended form. Each member of the high-activity class exhibits enhanced electrostatic complementarity as measured by visual examination of the residual potential. Although the strongest patches of noncomplementarity remain, the background of negative residual potential, which fills the TEM1 binding site of BLIP, is reduced. The residual potentials for TEM binding of the BLIP mutants D163K and D133K are shown in Figure 2.1E and G. When the three highest-activity mutations were combined, the computed effects were almost completely additive. This is consistent with the picture that adding just enough positive charge in appropriate locations on the surface of BLIP works to partially cancel the overly negative residual potential, but adding too much positive charge in one region overcancels the negative residual potential and makes it positive. Taken together, these results suggest that changing overall molecular charge alone is insufficient to improve binding affinity in this "electrostatically unbalanced" complex, but that when applied in the appropriate regions, increases in positive charge density can lead to computed enhancements in binding affinity.

A number of the mutations studied computationally here were also made and studied experimentally and computationally by Selzer and co-workers [89]. For these single mutations the qualitative agreement between experiment and computation is excellent. The low-activity class of mutation, with predicted electrostatic binding enhancements of 0.5 kcal/mol or less, all produced small binding enhancements experimentally (less than 1 kcal/mol, relative to wild type). The only single mutation for which our computations predict high activity that was studied, D163K, was shown experimentally to enhance binding affinity by 2.0 kcal/mol. A D163A mutant was also studied experimentally, and calculations made for this mutant show similar results (see Table 2.1). Four multiple mutants included in the study by Selzer *et al.*[89]

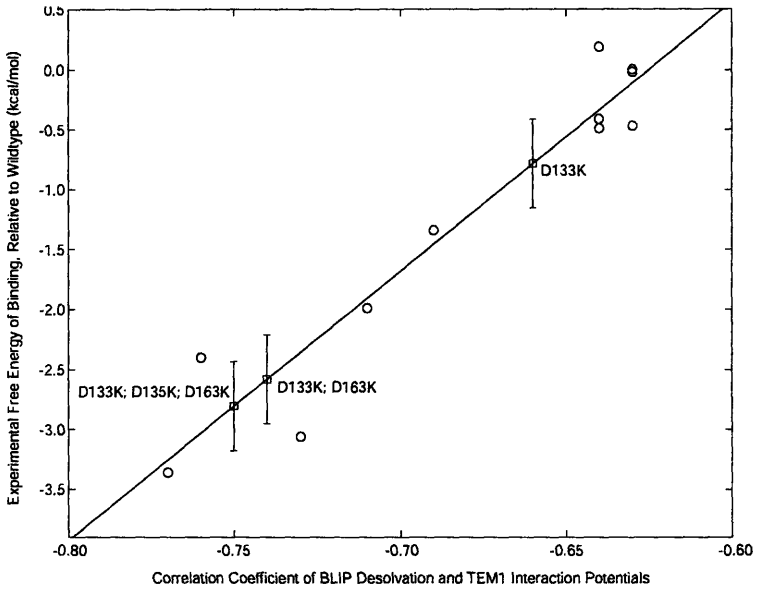


Figure 2.2: **Variation of experimental binding free energies with R**, the correlation coefficient between the BLIP desolvation potential and the TEM1 interaction potential, calculated on the surface of BLIP. Squares indicate mutants previously characterized [89]. The correlation coefficient of R and the experimental binding free energy is 0.96. The diagonal line indicates the least-squares best-fit. Circles indicate novel mutants characterized computationally, and are placed on the best fit line. Error bars on novel mutants indicate the standard deviation of points for which an experimental binding affinity has been calculated from the best-fit line.

---

were modeled and subjected to the same computational analysis. Again, the computational analysis reproduces the experimental division between a single low-activity mutant and three high-activity mutants. Overall, the calculated results show good agreement with the experimental data. One illustration of this is the strong correlation of the experimental binding free energies to those calculated here. The fact that similarly strong correlation is found between experimentally determined binding free energy and the residual potential statistic, R, (Fig. 2.2) suggests that electrostatic effects are a primary means by which these mutants act, as reflected in an improvement of overall electrostatic complementarity. Thus the residual potential and its quantitative analysis show significant promise as tools for understanding, and potentially designing, these types of surface mutations, some of which act via through-solvent interactions, to promote binding.

Three of the multiple mutants studied by Selzer et al. [89] contained the D163K mutation: N89K/D163K/V165K, V134K/D135K/D163K and N89K/D135K/D163K/V165K. These multiple mutants alter the total charge on BLIP by $+4e$, $+5e$, and $+6e$, respectively, and we calculate for them improvements in binding free energy of -9.3, -9.0, and -9.8 kcal/mol, respectively, relative to wild-type. Two of these contain the additional high-activity mutation D135K. The correlation of the interaction and desolvation potentials is improved for these mutants relative to D163K, and a decrease in the excess negative residual potential can be seen (see residual potential for the $+6e$ mutant, Figure 2.1F). These results also agree with experiment; N89K/D163K/V165K, V134K/D135K/D163K and N89K/D135K/D163K/V165K have experimental improvements in binding free energy of 2.4, 3.1, and 3.4 kcal/mol, respectively, all higher affinity than D163K alone. Despite the fact that each of these multiple mutants has a significant steric improvement in computed binding affinity, this improvement is prompted by the D163K mutation in every case. The strong correlation of the residual potential statistic, R, with experimentally determined binding free energies for these multiple mutants leads us to believe that electrostatics is nonetheless the primary cause for improvement with these mutations, and that the high computed steric improvement is the source of most of the discrepancy between experimental and computed values of $\Delta\Delta G_{bind}$ (see Figure 2.2).

The calculations suggest that the best previously uncharacterized mutant, D133K, could likewise be improved by combination with other favorable mutations. The multiple mutants D133K/D163K and D133K/D135K/D163K alter the net charge on BLIP by $+4e$ and $+6e$. For these structures, our calculations predict binding free energies of -9.4 and -9.9 kcal/mol relative to wild-type. It is noteworthy that the D133K/D135K/D163K triple mutant has a better computed binding free energy than any other mutant considered in this study. The residual potential for this triple mutant is shown as Figure 2.1I.

We have examined the computed binding free energy and electrostatic complementarity of a series of mutants of the β-lactamase inhibitor protein and analyzed the

results with comparison to experimental binding free energies to TEM1 β-lactamase. We find that the correlation coefficient of the BLIP desolvation potential and the TEM1 interaction potential on the surface of BLIP is strongly correlated to the experimental binding free energies. In addition, this increased correlation can be seen visually as a reduced residual potential in many cases. A previously uncharacterized mutation of Asp 133 to Lys is proposed, which calculations suggest would enhance binding affinity both alone and in concert with previously identified mutations. The effects of these mutations are localized to the extent that they act on patches of the surface, somewhat locally improving the residual potential. However, the interactions are not specific; three of the four most effective mutation locations (D133, D135, and Q161) are more than 7 Å from TEM1, and the D163K mutation has similar computed effects even in different conformations. This helps to confirm the overall mechanism by which these mutations act; relatively long-range electrostatic interactions act through a region of solvent to improve the overall electrostatic complementarity of the ligand for its target receptor. More generally, favorable action-at-a-distance electrostatic interactions may occur at regions of the protein surface that are close enough to the binding site to allow for a significant charge–charge attraction between ligand and receptor, but far enough away that the desolvation penalty incurred by placement of the charge is small. We expect that the action-at-a-distance interaction is used widely in biology, both as a means of improving binding when tight binding is required, and as a more general means of modulating free energy of binding to achieve a desired degree of affinity. Further work investigating the design of surface mutations that permute the residual potentials toward increased complementarity is on-going.

2.3 Materials and Methods

All calculations were performed using the 1.7 Å crystal structure of the BLIP–TEM1 complex solved by James and co-workers as an initial model [134]. Hydrogen atoms were added using the HBUILD facility [138] within the CHARMM computer program [139] with the PARAM22 all-atom parameter set [140]. Visual analysis of the hydrogen-

bonding patterns including ionizable groups indicated that all ionizable residues should retain their standard protonation states (structure solved at pH 8.8). This resulted in a net charge of $-2e$ for BLIP and $-7e$ for TEM1. Moreover, all water molecules were removed in the calculations (there was no interfacial solvent).

Residues on BLIP were selected for mutation by visual examination of all residues between 6 and 15 Å from any atom in TEM1 in the bound complex that also expose more than 40 $Å^2$ of solvent accessible surface area. From these, proline, cysteine, and glycine residues were discarded, as were any residues that appeared to make structurally significant intramolecular hydrogen bonds. In addition, we chose representative positions to mutate from strings of positions contiguous in sequence space. Finally, we chose to model the mutations to D163 suggested by Selzer et al. [89], and to D133 based on a continuum electrostatic analysis of the detailed contributions of the individual side chains of BLIP to TEM1 binding (DFG, BAJ, & BT, in preparation).

Model structures of single mutants to BLIP were generated by holding all backbone atoms and all non-mutated side-chains fixed, while allowing mutated side chains to take the lowest-energy conformation achieved by minimizing in CHARMM with a distance-dependent dielectric of $4r$ from seed locations generated by combinatorially scanning all side chain dihedral angles in 30° increments in the TEM1-bound state. Multiple mutant model structures were created by combining independently generated single mutant side chains when the mutations were located more than two residues apart in the BLIP sequence. When mutant residues were in closer proximity, the side-chain structures were generated simultaneously in the same manner as single mutants, but with coarser dihedral scanning in 120° increments.

Binding free energies were calculated as the sum of van der Waals, solvent-accessible surface area, and continuum electrostatic terms, using the approximation of rigid-body docking. The van der Waals contribution to binding free energy was calculated with the PARAM22 set of parameters for the program CHARMM. The solvent-accessible surface area contribution was calculated in the manner suggested by Sitkoff *et al.* [141], with the surface area contribution to the free energy of a struc-

ture calculated as 5.4 calories per square Ångstrom of surface area plus a constant of 920 calories. The contribution of burying surface area to binding free energy is then calculated as the difference between the complex free energy and the sum of the free energies contributed by the unbound BLIP and TEM1 surface areas.

Continuum electrostatic calculations were performed by numerical solution of the Poisson–Boltzmann equation, using a locally modified version of the program DELPHI [142–144] with PARSE atomic radii and partial atomic charges [141]. A grid of 257 × 257 × 257, with a spacing of 0.29Å, was used to calculate electrostatic binding free energy. Residual potentials were calculated from a coarser 129 × 129 × 129 grid to decrease the difficulty of storing and plotting surface potentials. For all electrostatic calculations, a protein dielectric constant of 4 and a solvent dielectric of 80 were used, along with an ionic strength of 0.145 M and a 2.0 Å ion exclusion layer. Surface potentials were displayed and numerically analyzed with locally developed software.

2.4 Acknowledgments

We thank Barry Honig for making DELPHI available, Martin Karplus for CHARMM, Michael Altman for in-house residual potential plotting software, and other members of our research group for helpful discussions. This work was supported by the National Institutes of Health (GM065418).

Chapter 3

A Computational Method for the Analysis and Prediction of Protein:Phosphopeptide-Binding Sites[1]

Abstract

Phosphopeptide-binding domains, including the FHA, SH2, WW, WD40, MH2, and Polo-box domains, as well as the 14-3-3 proteins, exert control functions in important processes such as cell growth, division, differentiation, and apoptosis. Structures and mechanisms of phosphopeptide binding are generally diverse, revealing few general principles. A computational method for analysis of phosphopeptide-binding domains was therefore developed to elucidate the physical and chemical nature of phosphopeptide binding, given this lack of structural similarity. The surfaces of nine phosphopeptide-binding proteins, representing seven distinct classes of phosphopeptide-binding modules, were discretized, and encoded with information about amino acid identity, surface curvature, and electrostatic potential at every point on the surface in order to identify local surface properties enriched in phosphoresidue contact sites. Cross-validation indicated that propensities corresponding to this enrichment calculated from a subset of the training data could be use to predict the phosphoresidue contact site on proteins not used in training with no false negative results, and with few unconfirmed positive predictions. The locations of phosphoresidue contact sites were then predicted on the surfaces of the checkpoint kinase Chk1 and the BRCA1 BRCT repeat domain, and these predictions are consistent with recent experimental evidence.

[1]This chapter has previously been published as:

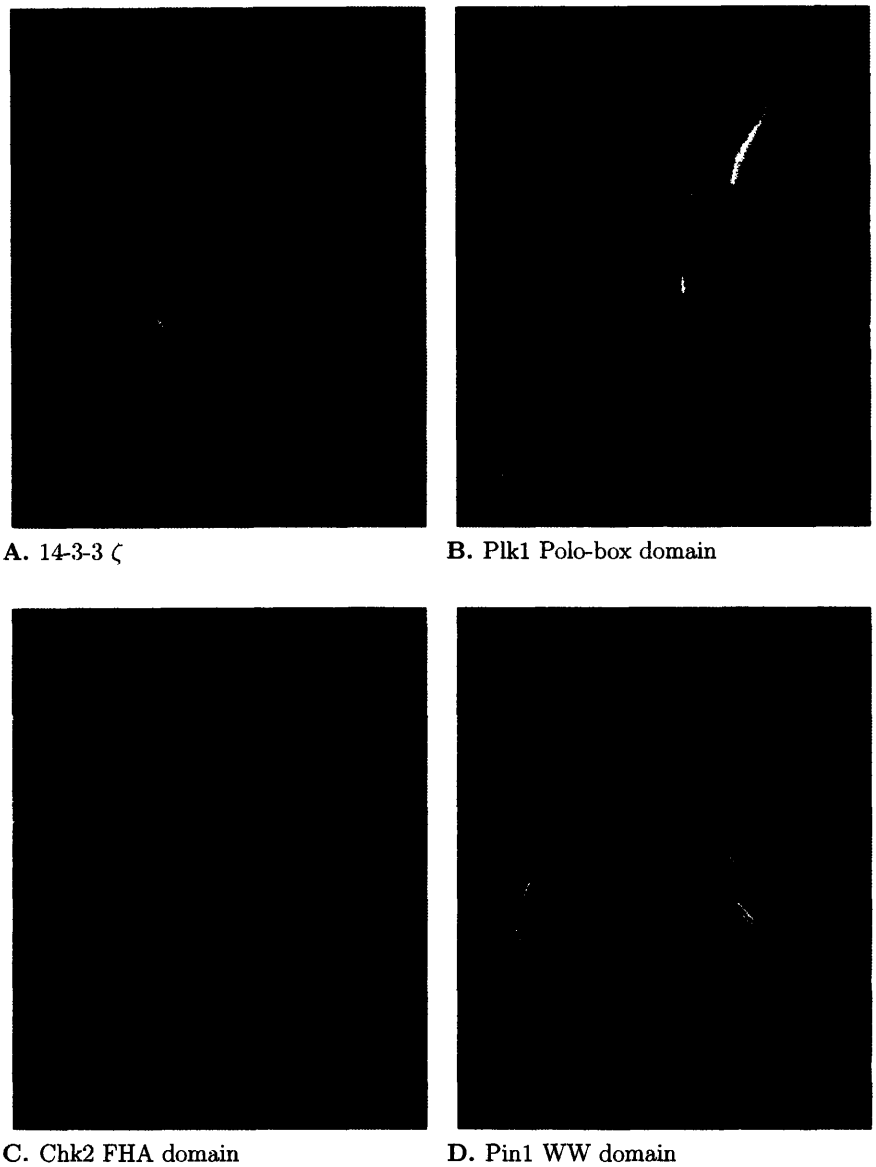
Brian A. Joughin, Bruce Tidor, and Michael B. Yaffe. A Computational Method for the Analysis and Prediction of Protein:Phosphopeptide-Binding Sites. *Protein Science*, 14:131–139, 2005.

3.1 Introduction

Many aspects of cellular biology, including cell cycle control, differentiation, and apoptosis, are regulated by the complex interplay of protein substrates with protein kinases, phosphatases, and phosphopeptide-binding domains [39–42]. Phosphopeptide-binding domains participate in signal transduction by recognizing and binding preferentially to the phosphorylated forms of specific proteins. In addition to binding directly to the phosphoserine, phosphothreonine, or phosphotyrosine residue, phosphopeptide-binding domains also recognize distinct linear sequence motifs surrounding the phospho-amino acid to achieve substrate specificity. To date, however, no comprehensive study has identified a unified set of physical-chemical, structural, or energetic requirements necessary and sufficient for phosphopeptide binding.

The structures of eight distinct classes of phosphopeptide-binding modules in complex with phosphorylated peptides or proteins have been solved (WW, PTB, SH2, MH2, FHA, WD40, Polo-box, 14-3-3). An examination of these structures reveals little structural similarity among the phosphopeptide-binding sites, apart from the evolutionary conservation seen among members of the same domain family [41] (see Figure 3.1). We reasoned that, despite the lack of gross structural similarity, there should be some underlying chemical and physical characteristics that define the phosphopeptide-interacting surface. We therefore analyzed a representative collection of these domains in detail and evaluated a set of physical and chemical properties at discrete points along their molecular surfaces. These properties were used to calculate a propensity value for each property to occur within a phosphoresidue contact site.

We found that these propensity values were able to correctly identify the phosphoresidue contact site on phosphopeptide-binding domains for which the site was known, in a cross-validation procedure. We used these propensities to predict the location of phosphopeptide-binding sites on the surface of two domains for which there was no published phosphopeptide co-crystal structure; the BRCT-repeat domain of the protein BRCA1, and the kinase domain of the checkpoint protein Chk1. BRCA1 is a tumor-suppressing protein whose dysfunction predisposes women to breast and



A. 14-3-3 ζ **B.** Plk1 Polo-box domain

C. Chk2 FHA domain **D.** Pin1 WW domain

Figure 3.1: **Structures of known phosphopeptide binding sites.** (A) 14-3-3 ζ[65]. (B) Plk1 Polo-box domain [64]. (C) Chk2 FHA domain [145]. (D) Pin1 WW domain [66]. Note that these proteins do not share a common secondary structure, nor do they make a common set of contacts with the phosphorylated side chain of their cognate peptide. This figure was rendered using the program RASTER3D [136].

ovarian cancer. The BRCT-repeat domains of BRCA1 and several other proteins were recently shown to bind phosphopeptides as part of the DNA damage response [58, 76]. The checkpoint kinase Chk1 plays a critical role in the cell cycle response to DNA damage, and appears to be regulated by binding to phosphopeptides at a site distinct from that of its catalytic activity [90]. The resulting predictions are corroborated with experimental data identifying the sites of phosphopeptide interaction. We anticipate that this computational approach to identifying phosphopeptide-binding sites will find general utility in the functional annotation of the structural genome, in the characterization of the structure and function of new phosphopeptide-binding domains as they are discovered, and in the identification of sites to target with inhibitors of protein/phosphopeptide interaction.

3.2 Results

To investigate the unifying principles involved in phosphopeptide recognition, we examined nine X-ray crystal structures representing seven phosphoserine-, phosphothreonine-, and phosphotyrosine-binding domains (Table 3.1). We observed little, if any, identity in the amino acids or their three dimensional arrangements within the phosphopeptide-binding sites [41]. Nevertheless, we felt that the physical and chemical requirements for phosphopeptide binding were in some manner encoded in these sites. We therefore built the phosphate-accessible molecular surfaces for each phosphopeptide-binding domain using a triangular mesh [146], and a probe radius of 3 Å, corresponding to the approximate radius of a phosphate ion. Each vertex on the mesh was encoded with information corresponding to a set of characteristics including amino acid identity, local mean surface curvature, and solvated electrostatic potential (see Materials and Methods). For each characteristic, the likelihood of occurrence at the contact sites for phosphorylated side chains was calculated. This likelihood was normalized by comparison with the likelihood of finding that same characteristic over the total phosphate-accessible surface area of the nine proteins studied, to derive a propensity for that characteristic being found in a phosphoresidue contact site.

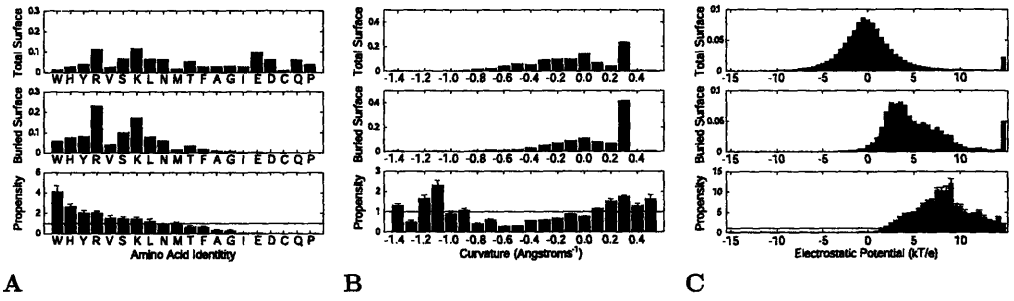


Figure 3.2: **Calculation of phosphoresidue contact propensities from global and phosphoresidue contact probability distributions.** Probability distributions over the total protein surface (upper panels), over the phosphoresidue contact surface (middle panels), and the phosphoresidue contact propensity (lower panels) were calculated for the properties (A) amino acid identity, (B) mean surface curvature, and (C) solvated electrostatic potential. Error bars in lower panels indicate twice the standard deviation of the mean for removing each crystal structure from the dataset, one at a time (N=9). The horizontal lines in the bottom panels indicate the mean phosphoresidue contact propensity, which is always equal to 1.

3.2.1 Phosphoresidue contact site properties

Amino acid identity

For the set of phosphopeptide-binding domains studied, the distribution of amino acids at all surface points unsurprisingly shows large contributions from charged amino acids, with arginine, lysine, and glutamic acid having the highest percentages observed (Fig. 3.2A, upper panel). In contrast, the highest percentages within the portion of the surface that contacts a phosphorylated side chain are contributed by arginine, lysine, serine, and tyrosine, while the acidic amino acids are almost never present (Fig. 3.2A, middle panel). Propensities for each amino acid to contact the phosphorylated serine, threonine or tyrosine side chains were calculated by normalizing the frequency of each amino acid at surface points in the phosphoresidue contact site by the frequency of that amino acid over the entire set of protein surfaces studied. This revealed the highest specific enrichment of tryptophan, histidine, tyrosine, and arginine, in that order, at phosphoresidue contact sites (Fig. 3.2A, lower panel).

While it might be expected that the positively charged amino acids lysine and

arginine would be the most over-represented in sites that bind negatively charged phosphates, this appears not to be the case, since lysine and arginine are extremely common on the surface of proteins in general, while tryptophan is not. While it would be quite surprising to find a phosphopeptide-binding site without lysine or arginine in it, the mere presence of a lysine or arginine on the surface of a protein carries less predictive weight than the presence of a tryptophan. There are three tryptophan residues in phosphoresidue contact sites in our data set, one on each of the proteins Pin1, Cdc4, and Plk1. In addition to contacting the phosphoresidue, all three tryptophans contact proline residues to the C-terminal side of the phosphoresidue of the phosphopeptide. This indicates a strong possibility that the high incidence of phosphoresidue-contacting tryptophans in our dataset may indicate the favorability of tryptophan/proline interaction in the context of the common phosphoresidue-proline motif. Interestingly, the contacts made between an arginine and a phosphorylated side chain typically involve a bidentate interaction with the guanadino group, while a tryptophan often stacks a large amount of its side chain surface against a phosphoresidue. Based on this observation, we independently calculated propensities for points on the surface of the three guanadino nitrogen atoms of the arginine side chain, and for the points on the remainder of the arginine residue. This revealed that the points associated with the nitrogen atoms have a high contact propensity, second only to that of tryptophan, while points on the rest of the amino acid are unlikely to be contacted (data not shown). This indicates that calculating propensities based on chemical functional groups, rather than amino acid identity *per se*, may serve to improve this analysis in the future, particularly once more structures are available from which to derive propensities. Several amino acids, including cysteine, glutamine, and proline were not observed to contact phosphorylated side chains, although this may be due to the relatively small size of the data set of known phosphopeptide-binding domain structures.

Surface curvature

A measure of the mean local curvature about each surface point was calculated [147], and used to produce a propensity value related to surface curvature. There is a spike in the overall distribution of surface curvatures at approximately 0.3 $Å^{-1}$, corresponding to the local concavity at any location where the 3 Å probe used to derive the molecular surface contacted three or more protein atoms (Fig. 3.2B, upper panel). There is also a small shoulder in the distribution centered at a convex curvature of -0.5 $Å^{-1}$, corresponding to regions where the probe touches only a single atom. The remainder of the distribution corresponds to saddle regions on the protein surface where the probe touches two atoms, and the surface has both concave and convex character.

Qualitatively, the distribution of surface points that bind to a phosphorylated side chain appears quite similar to the global distribution (Fig. 3.2B, middle panel). Quantitatively, however, the propensity for phosphoresidue contact, obtained by dividing the phosphoresidue contact site frequency distribution by the overall frequency distribution, is enriched in two regions (Fig. 3.2B, lower panel). One of these regions, with relatively high negative curvature values, is the ratio of sparsely populated regions of the contact site and global frequency distributions (Fig. 3.2B, upper and middle panels), making the predictive validity of propensities in this region questionable. The second region of high propensity lies between curvature values of 0.1 and 0.6 $Å^{-1}$ (Fig. 3.2B, lower panel), and corresponds to regions of concavity in the protein surface that are highly populated in the global distribution. The data in this region quantifies the well accepted tendency of ligands to bind to concave regions of protein surface, in the specific context of phosphopeptide-binding domain/ligand interactions.

Electrostatic potential

To examine the effect of electrostatic potential on phosphopeptide binding, we used a continuum electrostatic model to calculate the solvated state potential of each phosphopeptide-binding domain in our dataset in the absence of the cognate phos-

phopeptide ligand. The distribution of potentials on the phosphate-accessible surfaces of all proteins studied was bell-shaped, and centered approximately at zero (Fig. 3.2C, upper panel). As expected, the distribution of electrostatic potentials for the subset of the domain surfaces that contact a phosphorylated side chain is significantly shifted toward positive values (Fig. 3.2C, middle panel). As a result, the propensity distribution over electrostatic potentials, calculated as the distribution of electrostatic potentials in phosphoresidue contact sites divided by the global distribution of electrostatic potentials, peaks in the range between +7 and +9 kT/e.

As might be expected, the propensity for binding to phosphorylated side chains trails off as the electrostatic potential at a surface point becomes more negative from this peak, falling to almost zero at neutral electrostatic potential. Interestingly, the propensity also falls off for surface points having the highest electrostatic potential. The implication, then, is that surface points with such high positive electrostatic potentials are not as well suited for binding phosphopeptides as points with more moderate potentials, despite the high negative charge of a phosphorylated amino acid side chain. This is likely due to the high energetic cost of desolvating a region of such extreme positive potential [128].

3.2.2 Predictive ability for known phosphoresidue contact sites

To determine whether the calculated propensities were unduly influenced by any single structure in the data set, a cross-validation procedure was used ("jack-knifing") in which each structure was individually removed, and the propensities recalculated. The nine resulting sets of propensities were quite similar (shown by error bars in Figure 3.2, lower panels), with individual propensity values in well populated regions of the distributions differing on average from those calculated for the full dataset by less than 10% in the case of surface curvatures and electrostatic potentials, and by less than 25% for amino acid identities.

Of the three independent propensities calculated for amino acid identity, surface

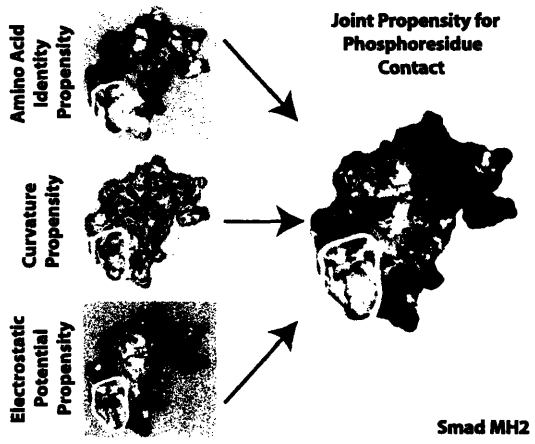


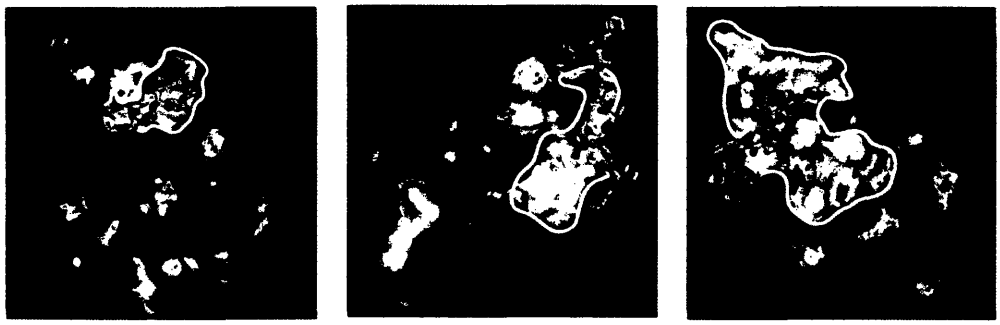
Figure 3.3: **Calculation of joint propensity for phosphoresidue contact.** Propensities were calculated independently for amino acid identity (upper left), local mean surface curvature (middle left), and solvated electrostatic potential (lower left). These propensities were combined multiplicatively to obtain a joint propensity for phosphoresidue contact (right). Two linear scales were used to depict unfavorable and favorable propensity. Unfavorable propensity values from 0 to 1 are colored from red to white. Favorable propensity values are colored from white to blue over the values 1 to 4 for amino acid identity, 1 to 2.5 for surface curvature, 1 to 12 for solvated electrostatic potential, and 1 to 20 for joint phosphoresidue contact propensity. In some regions, as with the area outlined in yellow, the three individually calculated propensities combine constructively to create a large region of favorable joint propensity. In other regions, such as the area outlined in green, an area that looks favorable for phosphoresidue contact by one measure, such as electrostatic potential, combines with the propensities generated by other characteristics to define a site that is less favorable, overall, for phosphoresidue contact.

---

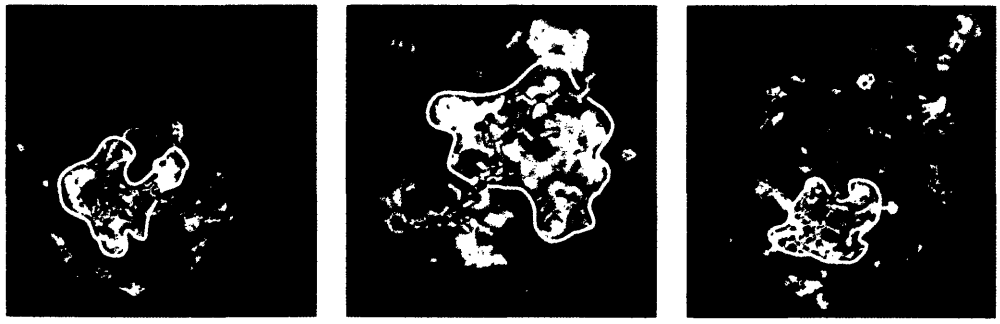
curvature, and electrostatic potential, none was sufficient on its own to unambiguously identify the site of known phosphoresidue contact on the set of phosphopeptide-binding domains studied here (Fig. 3.3, left panels). However, the scales of propensities encountered in this analysis provide a framework for understanding the contribution of each characteristic studied to phosphoresidue binding. The scales of propensity values encountered indicate the most favorable values of electrostatic potential are more predictive, with respect to phosphoresidue contact, than the most favorable values of amino acid identity or surface curvature. Nevertheless, unfavorable propensity values contributed by amino acid identity or surface curvature are capable of countering false positive favorable contributions from positive electrostatic potential in order

to improve the accuracy of our predictions, as shown in Figure 3.3.

We next investigated whether the amino acid identity, surface curvature, and electrostatic potential propensities could be combined in a prospective manner to identify phosphoresidue contact sites (see Figure 3.3, Materials and Methods) using cross-validation. For each structure in our set of known phosphopeptide-binding domains, the joint propensities calculated from every other member of the set were painted onto the surface of the domain of interest and visually inspected. As shown in Figure 3.4, the correct phosphate binding site was easily identified in every case as a contiguous region of mixed high and neutral joint propensity. No false negative prediction of a phosphoresidue contact site was made. In most cases, including that of the protein 14-3-3 ζ, the Smad MH2 domain, and both FHA and both SH2 domains studied, only a single site of significant size and propensity was observed. However, for Pin1, Cdc4, and the Polo-box domain of Plk1, a second site of comparable size and propensity to a known real phosphopeptide-binding site was also observed (see Figure 3.5). Intriguingly, the second predicted phosphoresidue contact region on the Pin1 surface lies at the catalytic site in Pin1's proline isomerase domain. This site is known to bind specifically to, and isomerize, phosphopeptides containing the same motif as that recognized by the WW domain [55], and therefore corresponds to a phosphopeptide-binding site. In the case of Cdc4 and the Polo-box domain of Plk1, the second predicted phosphopeptide-binding site may represent false positive predictions, or may indicate sites of further interaction with as-yet-unidentified phosphopeptides or other anionic ligands. It is also of interest to note that in most cases, the region of favorable propensity detected is quite a bit larger than the sites of phosphoresidue contact on which the method was trained. This indicates that the local properties most enriched in sites of phosphoresidue contact are also highly enriched in the surrounding regions. This may be reflective of a kinetic mechanism for attracting the phosphopeptide ligand to its binding site.



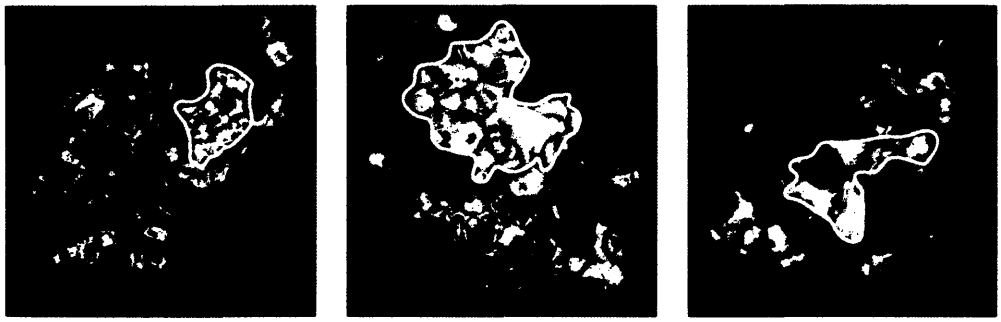
A. Pin1 WW B. Rad53 FHA1 C. Chk2 FHA



D. Smad MH2 E. p56 SH2 F. Cdc4 WD40

G. 14-3-3 ζ H. Src SH2 I. Plk1 Polo-box

Figure 3.4: **Cross-validation of phosphoresidue contact site predictions on known phosphopeptide binding domains.** Phosphopeptides are shown in a licorice representation, with the phosphate phosphorus atom colored green. Surface coloring is linear from red to white for unfavorable propensity values from 0 to 1, and from white to blue for favorable propensity values of 1 to 30. Predicted phosphoresidue contact sites are outlined in yellow. The phosphopeptide-binding domain shown is indicated within each panel. The skinny sticks in panel (D) indicate the Smad monomer that contains the bound phosphopeptide. The phoshotyrosine phosphates in panels (E) and (H) are buried beneath the phosphate accessible surface.



A. Cdc4 B. Plk1 C. Pin1

Figure 3.5: **Additional phosphoresidue contact site predictions.** Additional site predictions were made on the surfaces of the indicated proteins. Surface coloring is linear from red to white for unfavorable propensity values of 0 to 1 and from white to blue for favorable propensity values from 1 to 30. Predicted phosphoresidue contact sites are outlined in yellow. The predictions for Cdc4 and Pin1 did not lie on the phosphopeptide-binding domains of those proteins.

3.2.3 Prediction of the phosphoresidue contact sites of Chk1 kinase and the BRCA1 BRCT-repeat domain

The method under development here is capable of predicting the location of phosphoresidue contact sites on the surface of phosphopeptide-binding domains whose unliganded structures are known. These predictions can then be investigated experimentally. Two such cases are currently available. The checkpoint kinase Chk1 has been found to be regulated by binding to the phosphorylated form of the protein claspin [90] at a site within the kinase domain. The BRCT-repeat domains of several proteins, including BRCA1 and PTIP [58, 76] have recently been identified as phosphopeptide-binding domains. One crystal structure of the Chk1 kinase domain, and three crystal structures of BRCA1 BRCT-repeat domains, in the absence of bound phosphopeptide are available. We therefore applied our method to these structures.

Application of local surface propensity analysis to the Chk1 kinase domain surface identified two possible sites for phosphopeptide binding (Fig. 3.6A). These sites are connected by a small region of neutral propensity. The first site, located at the inter-

A. Chk1 kinase domain **B.** BRCA1 BRCT

Figure 3.6: **Predicted phosphoresidue contact sites on the surfaces of Chk1 and BRCA1.** Surface phosphoresidue contact propensity plots (top panels) and secondary structure (bottom panels) with residues named in the text shown in licorice. (A) Chk1 kinase domain. On the surface plot, site one is outlined in yellow on the right, and site two is outlined in yellow on the left. In the secondary structure diagram, the small lobe of the Chk1 kinase domain is colored blue, and the large lobe is colored red. (B) BRCA1 BRCT repeat domain. The left panel indicates the first predicted site, which has been shown experimentally to be the site of phosphopeptide binding [78, 91, 92], and the right panel indicates the second predicted site. The axis shown indicates the axis of rotation between the shown molecular faces. In the secondary structure diagrams, the first BRCT repeat, the linker, and the second repeat are colored blue, green, and red, respectively. Surface coloring is linear from red to white over the unfavorable propensity values 0 to 1 and from white to blue over favorable propensity values from 1 to 30. Portions of this figure were generated using the programs MOLSCRIPT [135] and RASTER3D [136].

face between the large and small lobes of the kinase, but not in the kinase catalytic site, is made up of the amino acid side chains K54, R129, T153, R162, and N165 (Fig. 3.6A, rightmost indicated site). The mutations K54A, R129A, and T153A, and R162A have all been shown to abrogate claspin binding in the frog Chk1 homolog Xchk1 [90]. Our results suggest that those residues are directly responsible for phosphoclaspin binding. The second site we identified, on the small lobe of the kinase domain, is adjacent to the first, and is made up of the Chk1 amino acid side chains K53, K60, H73 and R75 (Fig. 3.6A, leftmost indicated site). While this site has not previously been identified as a site of phosphopeptide binding, it is known that phosphoclaspin binding to Xchk1 requires two separate claspin phosphorylation events, on residues S864 and S895. It is possible, therefore, that the two phosphopeptide residues pS864 and pS895, separated by 31 amino acids, are recognized by two distinct phosphopeptide-binding sites on the Chk1 surface.

Two predicted phosphopeptide-binding sites were also identified on the surface of the rat BRCA1 BRCT-repeat domain. The first of these is a bowl-shaped depression entirely within the first of the two BRCT repeats in the structure. The surface that composes the site is contributed by three amino acid side chains – K1648, S1601, and T1646 (Fig. 3.6B, upper panel). This triad of residues is conserved in the BRCA1 protein of humans. The other potential binding site is found in a channel composed of four amino acids at the interface between the second BRCT repeat and the helix linking the two repeats – R1697, R1791, H1692, and R1793 (Fig. 3.6B, lower panel). R1697 and H1692 are conserved in humans, while R1791 and R1793 are both glutamine in human BRCA1. Thus, the method presents two hypotheses for the site responsible for phosphopeptide-binding activity, which are readily tested by site-directed mutagenesis experiments.

During the preparation of this manuscript, the crystal structure of the human BRCA1 BRCT domain in complex with a phosphopeptide was solved [78, 91, 92]. In this structure, the phosphoresidue contact site was shown to correspond to the first of the two sites on the BRCA1 surface predicted by our method, indicating that for this site at least, our prediction was correct. This result, together with the experimentally

corroborated prediction on the surface of the Chk1 kinase domain, indicates that the methodology described here has captured a large portion of the chemical and physical nature of phosphopeptide binding in a manner that is useful for predicting binding sites.

The phosphoresidue contact site predictions described here were originally made by visual inspection of the joint phosphoresidue contact potential on the surfaces of Chk1 and BRCA1 and selection of the largest site of favorable propensity. We are currently exploring a vertex-clustering algorithm designed to identify large regions of favorable propensity in an automated fashion.

3.3 Discussion

We have developed a novel framework for phosphopeptide-binding site prediction. Our method is based on finely discretizing the surface of proteins, identifying physical and chemical properties that are over-represented on those surfaces at sites of contact with phosphorylated amino acid side chains, and locating contiguous patches of those properties on the surfaces of proteins for which a prediction is to be made. Previous methods for the discovery of functional sites on proteins include patch analysis [148, 149], in which properties are calculated for a number of large overlapping surface patches, and used in conjunction with heuristics to identify functional sites; and evolutionary trace analysis [150], which depends on a large number of homologous protein sequences to find clusters of evolutionarily conserved residues. In contrast, the method described here, which uses discretized surface propensities, is capable of using a relatively small number of structures to determine local surface properties enriched in a functional site. The local nature of the surface properties analyzed appears to capture some of the physical and chemical properties required for phosphopeptide binding, despite the larger-scale dissimilarity of the binding sites used in training.

There are three important caveats to the computational method. First, we assume the independence of propensities calculated from a set of properties – amino acid identity, mean surface curvature, and electrostatic potential – which are not

themselves independent. In the limit of a large volume of data, it is possible to abandon this approximation by calculating an exact propensity value for every possible combination of property values. As more data become available, it should be possible to learn correct parameters for the combination of these propensity values.

Sites with the highest propensities for phosphoresidue contact have strong favorable propensity contributions from each of the three properties considered here. In the limit of currently available data, we find that all three properties considered here are necessary for accurate site prediction. Although strong favorable propensity for phosphoresidue contact is driven by the solvated electrostatic potential, false positive predictions that would be generated by the consideration of electrostatics alone are avoided by combining information about surface curvature and amino acid identity.

Second, we calculate and cross-validate propensity values from a set of crystal structures solved in the presence of phosphopeptide. These structures may involve some induced fit to their cognate peptides, whereas structures for which useful predictions can be made would be in their unliganded *apo* conformation. Despite this, we make predictions for the Chk1 kinase domain and the BRCA1 BRCT-repeat domain that are validated by experiment, indicating that the physical and chemical aspects of a phosphoresidue contact site that are captured by our model are not lost in the *apo* state.

Finally, the method described here is designed to identify the site of phosphoresidue contact on the surface of a known phosphopeptide-binding domain. It is clear that as novel phosphopeptide-binding domains are discovered, and as structural genomics efforts come to fruition, this approach will prove useful in rapidly identifying the functional sites on unliganded crystal structures without necessitating further crystallographic effort. Because the propensities calculated here are trained to differentiate phosphoresidue contact surface from the remainder of the surface of phosphopeptide-binding domains, this may be less useful in mining structural databases for novel phosphopeptide-binding domains. We expect, based on the emphasis given by our propensity scale to positive electrostatic potential, that this scale might score some anion- and phosphate-binding sites quite favorably. This has been con-

firmed by our examination of several nonphosphopeptide-binding proteins (data not shown). However, if the goal of future work is to differentiate among different types of anion-binding sites, appropriate propensity scales and other machine learning tools could certainly be developed, for example for the differentiation of phosphoresidue contact sites from such "decoy" sites.

The method described here is highly extendable, both in terms of the type of functional site examined, and in the characteristics for which propensities are calculated. Propensity calculations can be performed on continuous properties such as curvature and electrostatic potential, which have been discretized via binning, as well as on traditional discrete properties such as amino acid identity. Therefore, any property that can be assigned to the vertices of a protein surface can be applied to site predictions within this methodological framework. Moreover, predictions can be made within this framework for any functional categorization for which predictive physical surface properties can be found. Our successes in the identification of phosphoresidue contact sites on the surfaces of the Chk1 kinase domain and the BRCA1 BRCT-repeat domains indicate the utility of this methodology in functional site annotation.

3.4 Materials and Methods

3.4.1 Structures

The structures used as a training set in this study (Table 3.1) were selected as being the best high-resolution crystal structures representative of the known phosphopeptide-binding domain/peptide interactions. Structures of one 14-3-3 protein, one group IV WW domain, one WD40 domain, one MH2 domain, two FHA domains, and two SH2 domains were used to gather propensity data. The single most well resolved structures of the Chk1 kinase domain (PDB Code 1IA8) [151] and BRCA1 BRCT-repeat domain, from the rat BRCA1 protein, (PDB Code 1L0B) [152] were used for phosphoresidue contact site predictions.

Table 3.1: **Structures used to calculate propensity data.**

PDB ID	Protein	Domain Type/ Phosphorylated AA	Surface Points	Ref.
1F8A	Pin1	WW / 2x pS	35,582	[66]
1G6G	Rad53	FHA / pT	24,372	[61]
1GXC	Chk2	FHA / pT	24,293	[145]
1KHX	Smad	MH2 / 2x pS	40,832	[63]
1LCJ	$p56^{Lck}$	SH2 / pY	21,905	[67]
1NEX	Cdc4	WD40 / pT	76,798	[153]
1QJB	14-3-3ζ	14-3-3 / pS	46,068	[65]
1SPS	Src	SH2 / pY	21,954	[154]
1UMW	Plk1	Polo-box / pT	40,426	[64]

3.4.2 Propensity calculation

For each property associated with a surface point – amino acid identity, surface curvature, and electrostatic potential – a propensity for phosphoresidue contact was calculated. The propensity of a property i was calculated as:

$$P(i) = \frac{n_b(i)/n_b}{n_t(i)/n_t}, \tag{3.1}$$

where $n_b(i)$and $n_t(i)$are the number of surface points with characteristic i contacting phosphoresidues and in total, respectively, and n_b and n_t are the number of surface points contacting phosphoresidues and total number of surface points in the data set, regardless of characteristic.

When attempting to predict the phosphoresidue contact site on a protein, the propensity assigned to each surface point was computed, under the simplifying assumption that propensities generated using amino acid identity, local mean surface curvature, and solvated electrostatic potential combine noncooperatively, as

$$P = P_{aa} \times P_{curv} \times P_{es}. \tag{3.2}$$

Figure 3.3 shows one example of the combination of these three individual propensities

to derive a joint propensity.

3.4.3 Surface and contact calculation

The program MSMS [146] was used to obtain a triangular surface mesh for each phosphopeptide-binding domain, using a probe radius of 3.0 Å, the approximate radius of a phosphate ion, and a surface density of 5.0 vertices/Å^2. Calculations were performed on a monomer of each phosphopeptide-binding domain in the presence and absence of only the phosphorylated side chain of the corresponding binding peptide. Surface points contacted by the phosphoresidue were identified as those that were surface accessible on the unliganded protein surface but buried in the protein/phosphoresidue complex surface such that it was further than 0.3 Å from the nearest point on the bound-state surface.

3.4.4 Amino acid identity assignment

The amino acid identity of each surface point was recorded as identified by MSMS, with points on the reentrant phosphate-accessible molecular surface assigned to the nearest atomic van der Waals sphere.

3.4.5 Mean surface curvature assignment

The mean surface curvature at each point was calculated according to the method of Meyer *et al.* [147]. In order to discretize the space of curvatures for propensity calculation, surface curvatures were binned with a bin width of 0.1 Å^{-1} between the values of -0.6 and 1.4 Å^{-1}, with curvatures above and below the extrema placed in the highest and lowest bin, respectively. Calculated propensities were found to be insensitive to the bin size selected over a range of bin sizes from 0.05 to 0.5 Å^{-1}

3.4.6 Solvated electrostatic potential assignment

The electrostatic potential at each surface point was calculated with a continuum electrostatic model with a locally modified version of the program DELPHI [142–144].

The calculation used the phosphopeptide-binding domain alone, a solvent dielectric of 80, a salt concentration of 0.145 M, a protein dielectric of 4, and PARSE parameters [141]. Prior to calculating potentials, hydrogen atom positions were added to the protein structures using the program REDUCE [155]. Electrostatic potentials were discretized for propensity calculation by binning, with bin with 0.5 kT/e, with data below -15 kT/e or above +15 kT/$e$assigned to the lowest and highest bin, respectively. Calculated propensities were found to be insensitive to the bin size selected over a range of bin sizes from 0.25 to 5.0 kT/e.

3.5 Acknowledgments

We would like to thank members of the Tidor and Yaffe laboratories for helpful discussion, and particularly Michael Altman for much of the development of our graphical analysis software. This work was partially supported by the National Institutes of Health (GM060594 to MBY and GM065418 to BT) and by a Burroughs-Wellcome Career Development Award to MBY.

Chapter 4

Computational Design of a Library of WW Domain Variants Targeting an Altered Ligand Specificity

Abstract

We have attempted to use both traditional single-sequence computational protein design and protein library design to identify WW domain variants capable of specifically binding peptides containing the motif phosphoserine-glutamine, "pS/pT-Q". Such peptides are potential products of phosphorylation by the DNA damage kinases ATM and ATR. Traditional protein design methods produced protein-peptide complex structures that looked, *in silico*, unlikely to be specific. However, we were able to design a library of protein sequences that might be able to bind "pS/pT-Q" peptides specifically by positioning amino acids to make multiple simultaneous hydrogen bonds to the peptidyl glutamine residue, while maintaining an intact phosphoserine binding site from the wild-type Pin1 WW domain. The Pin1 WW domain binds peptides containing the motif "pS/pT-P". This library is over 30-fold smaller than the corresponding combinatorial protein library, but contains all sequences that were computationally predicted to have the potential to make three hydrogen bonds to a peptidyl glutamine. It is likely, particularly in designs involving more than the five amino acid positions considered in this design, that focusing on noncombinatorial regions of sequence space enriched in a function of interest will prove useful in increasing the odds of success in experimental screening.

4.1 Introduction

The kinases ATM, ATR, and p38 are responsible for initiating the downstream response to DNA damage [93, 94, 156]. ATM and ATR share the same ligand phosphorylation motif; they tend to phosphorylate their ligands on a serine or threonine residue that is followed by a glutamine residue ("S/T-Q"), with particular affinity for ligands containing the sequence "L-S-Q-E" [157, 158]. Many important *in vivo* ATM and ATR ligands have been identified as being phosphorylated as a response to DNA damage, including Chk1, Chk2 [159], BRCA1 [160, 161], p53 [162–166], and others. It is unlikely, however, that all ATM and ATR ligands, or even all ATM and ATR ligands with clinical importance, have been identified.

We decided, therefore, to attempt the design and development of a laboratory reagent that would specifically bind to peptides and proteins containing the motif "pS/pT-Q". Such a reagent would be useful in the identification in cell lysates of potential products of ATM and ATR phosphorylation. Several other techniques are available for the identification of kinase targets [167]. Of these, the use of a designed reagent is conceptually most similar to the use of phosphomotif-directed antibodies [24]. Because phosphomotif-directed antibodies are raised in animals against degenerate libraries containing the motif of interest, it is difficult to ensure that the antibody binds specifically to all of the motif elements of interest. The development of a framework for the computational design of motif-specific affinity reagents would allow for specification by the researcher of the important determinants of specificity.

Successful instances of computational protein design [122–125, 168] have generally phrased design as an inverse of the protein folding problem [97, 98]: given a protein backbone structure of interest, identify the amino acid sequence that best stabilizes it. The analogous problem in the design of peptide binders is the identification of protein sequences that best stabilize a target complex structure relative to the separated components, given backbone structures for both ligand and binder. Specificity design has directly been considered a number of times. Design strategies have ranged from simple affinity optimization of the desired complex, without an explicit consideration

of specificity [169–171], to the incorporation of negative design elements to prevent interactions of the with wild-type molecules [172–177].

In designing a novel protein-ligand complex, rather than stabilizing an existing protein or altering the affinity of an existing complex, it is not obvious what existing complex or structure to use as a model for computation. Two feasible choices here are the ATM or ATR kinase domains themselves, which already bind to the "pS/pT-Q" ligand of interest after catalyzing its formation but before product release. In a sense however, these domains have evolved to release their product, and it is not clear that it would be straightforward to design a catalytically inert kinase domain variant that would bind phosphopeptides more tightly than its original peptide and ATP substrates. Moreover, there is no crystal structure of ATM or ATR in complex with its "S/T-Q" substrate to use as a model. We chose instead to alter the specificity of an existing phosphopeptide-binding domain.

Phosphopeptide domains operate in many major biological processes, including control of the cell cycle and cell fate determination [39–42]. Of the known phosphopeptide binding domains, we selected the group IV WW domain [54] as the strongest candidate for use as a design scaffold. Group IV WW domains bind to targets on the basis of a "pS/pT-P" motif, which matches the pattern we wish to bind – a phosphorylated residue, and one residue neighboring it on the C-terminal side. On this basis, we selected the structure of the Pin1 WW domain in complex with a "pS-P" peptide (PDB code 1F8A [66]) as the scaffold for our design.

A preliminary attempt to design individual WW domain variants with high affinity for "pS/pT-Q" gave disappointing results, leading us to develop instead a large library of possible binders to be screened experimentally. Computational design procedures have traditionally been used to generate a small number of protein sequences for individual testing. More recently, however, researchers have begun applying similar techniques to the design of protein libraries [178–180]. Computational design of protein libraries has generally been focused on the development of combinatorial libraries, which are more easily expressed and screened than irregularly shaped protein spaces. Here we attempt simply to generate a list of the protein sequences most likely

to bind "pS/pT-Q" ligands, without requiring that the resultant library be combinatorial. In conjunction with the oligonucleotide library design method presented in Chapter 5, we expect noncombinatorial libraries to be a powerful way to focus an experimental library screen on the portions of sequence space most likely to contain the protein function of interest.

4.2 Materials and Methods

4.2.1 Structure Preparation

An all-atom model of the Pin1 WW domain in complex with a "pS-P" peptide was generated, using a 1.84 Å resolution crystal structure [66] as a basis. The proline isomerase domain (residues 55-167) was removed. Density was missing from Pin1 residue 43 to 54, leaving only the WW domain (residues 1-42) and its ligand. Upon visual examination of hydrogren-bonding patterns, histidine residues 3 and 31 were designated as being neutrally charged and protonated on the ϵ nitrogen. The well-coordinated phosphoserine residue of the ligand peptide (174) was left in the doubly anionic deprotonated state based on an examination of the hydrogen-bonding pattern of the residue, the strongly cationic environment, and the fact that the pK_a of methyl phosphate at biological temperature is about 6.6 [181], only slightly more acidic than the cellular pH. The more solvent-exposed phosphoserine (residue 171) was trimmed back to serine after preliminary calculations indicated that the excess negative charge of the second phosphoserine residue biased the results of design calculations in favor of cationic WW variants, and this phosphoserine will not be present in most biological ligands of interest. The rotation about the terminal χ angle of asparagine, glutamine, and histidine residues was examined, but left in the state chosen by the crystallographer. We switched the coordinates of the carbonyl (C,O) and side-chain (CB, OG) atoms of the C-terminal residue of the WW-domain itself, serine 42, as the crystallographic designations left no room for a subsequent residue to be placed without a steric clash. Hydrogens were positioned using the HBUILD functionality

[138] of the program CHARMM [139] using the PARAM22 parameter set [140] with a constant dielectric of 1. The WW domain and peptide ligand were patched with protonated N-terminal amino groups. The WW domain was patched with a C-terminal N-methylamide, as the residues following 42 were missing crystallographic density. The phosphopeptide C terminus was also patched with N-methylamide, to avoid biasing the design with an excess of negative charge on the peptide.

4.2.2 Individual Protein Design

A design of individual WW-domains in stable complexes with a "pS-Q" ligand was undertaken, using a two-level treatment of electrostatics and solvation [182–185] performed as described in [176]. Briefly, nonpolar hydrogens were removed from the model structure generated according to the method given above. The peptide ligand residue 175 (a proline directly to the C-terminal side of phosphoserine 174) was mutated to glutamine and allowed to adopt any conformation in the Dunbrack and Karplus backbone-independent rotamer library (May 2002) [186, 187], supplemented with additional rotamers at $\pm$ 10° at χ_1 and χ_2. Notably the ϕ and ψ backbone dihedrals of the mutated proline are accessible to the amino acid glutamine, as well. Residue 176 was truncated to alanine to leave room for coordination of the glutamine by the WW domain. On the WW domain itself, five amino acid positions (23, 25, 27, 36, and 38) were selected as being capable of reaching the peptidyl glutamine while leaving the phosphoserine binding site essentially intact, and the five positions were given the ability to mutate to any of the natural amino acids excluding proline using the same rotamer library as the peptidyl glutamine. Notably, residue B38 is a tryptophan, one of the two "W" residues that gives the WW domain its name. In the wild-type Pin1 WW domain structure, this tryptophan stacks against the proline residue of the "pS-P" ligand, and we hypothesize that it is not required for the folding or binding activity of a "pS-Q"-binding variant.

A locally-authored version [184] of the DEE/A* algorithm [99, 101–103, 106, 107] was used to identify all protein sequences with an approximate sequence energy within 30 kcal/mol of the global minimum sequence energy using a pairwise energy

function consisting of internal geometry terms, a van der Waals term, and Coulombic electrostatics with a $4r$ dielectric, as calculated by CHARMM [139] using the PARAM19 parameter set [188, 189]. A model compound consisting of separated single side chains with neutrally-blocked termini was used to represent the unfolded protein state. Because we had no explicit decoy state that we wanted to develop specificity against, we optimized on the stability of the protein-"pS-Q" complex, which is calculated as the difference in energies between the bound-state energy and the unfolded model compound. The approximation of the global minimum stability of each sequence considered is made using the method of Mendes et al. [105], treating all conformations of an amino acid at a position as a flexible rotamer family. For the 20,000 sequences with the best sequence energy (a range of about 27 kcal/mol), the 10 lowest-energy structures were generated that varied to the extent that no two structures had all residues in the same cluster of $\pm$ 10° of χ_1/χ_2 sampling. In this step, no sequence approximation is used; the 10 structures generated have exact energies calculated, though by the inaccurate pairwise function described above. The lowest of these is termed the "low-accuracy" energy of the protein sequence.

All of these structures were re-evaluated by replacing the $4r$-dielectric Coulombic electrostatics with linearized Poisson–Boltzmann electrostatics and a SASA-dependent hydrophobic term of 5 cal/(mol Å^2) [141]. The linearized Poisson–Boltzmann equation was solved using a locally-modified version of the program DELPHI [142, 143, 190]. The linearized Poisson–Boltzmann solution was performed by placing the WW-domain/peptide complex on a 129×129×129 grid, filling first 23% of the grid using Coulombic electrostatics with Debye–Hückel screening at the external dielectric constant to fix boundary conditions, and then 92% of the grid using the 23%-fill solve to set boundary conditions. Protein dielectric was set at 4, solvent dielectric at 80, and salt was set to a concentration of 145 mM. Atomic radii and partial atomic charges were set according to the PARSE parameter set [141]. The molecular surface, using a probe radius of 1.4 Å to define the dielectric boundary, and an ion-exclusion layer of 2 Å, were used. The lowest energy, calculated over the 10 structures used for each protein sequence, is termed the "high-accuracy" energy

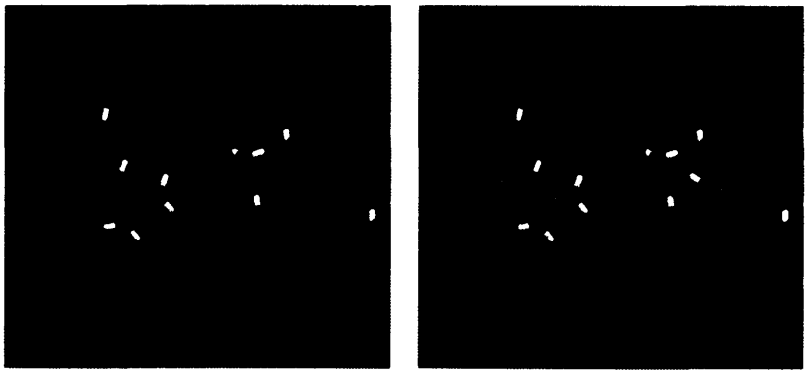
of the sequence.

4.2.3 Protein Library Design

It was our intuition, guided by previous studies of the role of electrostatics in protein-protein interaction [172, 191–193] and the coordination by several hydrogen-bonding groups of glutamine by the glutamine binding protein [194] that any protein with specific glutamine binding would achieve that specificity through hydrogen bonding. The procedure by which a WW domain variant library was designed is summarized in Figure 4.1. The wild-type Pin1 WW domain as prepared above was used as an initial model (Figure 4.1A). WW domain side chains 23, 25, 27, 36, and 38 and peptidyl residue 175 were removed from the structure (Figure 4.1B). The Dunbrack and Karplus backbone-independent rotamer library (May 2002), [186, 187], supplemented with additional rotamers at $\pm$ 10^o at χ_1 and χ_2, was used to identify every way of placing a side chain on the WW domain design positions and a glutamine side chain on peptidyl residue 175 such that a hydrogen bond was made and no steric clashes were generated (Figure 4.1C). We used a very permissive definition of a hydrogen bond; any two side chains that put a donor group hydrogen within 2.4 Å of a receptor atom on the opposite residue was kept. We considered as clashing any rotamer that had an interaction energy of greater than 25 kcal/mol with the fixed portion of the protein, or any pair of rotamers that had an interaction energy of greater than 25 kcal/mol with each other. We then built up sequentially from this data all ways of placing a peptidyl glutamine rotamer with 2, and then 3 rotamers on the WW domain such that all WW domain residues made simultaneous hydrogen bonds to the glutamine, and no residue or pair of residues clashed (Figure 4.1D). This produced a list of 487,005 structures.

The results of this procedure were evaluated by eye, in conjunction in some cases with minimization using the program CHARMM [139] with the PARAM22 parameter set [140] and an ACE implicit solvent model [195–197]. The goal of this evaluation was to find structures that might, with backbone relaxation and the ability to fill the two design positions not-yet-filled with any amino acid, feasibly form three simultaneous

A. Wild-type Pin1 WW domain **B.** Side chains removed



C. Single hydrogen bond **D.** Triple hydrogen bonds

Figure 4.1: **Library Design Method.** In all images, the WW domain is blue and the peptide is red. WW mutable residues 23, 25, 27, 36, and 38 are shown in licorice, as is the peptidyl residue 175. (A) The wild-type Pin1 WW domain in complex with a "pS-P" peptide (red). The peptidyl proline stacks between a tryptophan and tyrosine. (B) WW side chains 23, 25, 27, 36, and 38 and peptidyl side chain 175 are removed. (C) All ways of placing a glutamine rotamer on the peptide and any rotamer on any WW domain position such that each residue has an energy of interaction with the fixed portion of the protein of less than 25 kcal/mol, the pair of residues has an interaction energy of less than 25 kcal/mol, and a hydrogen bond donor from one residue is placed within 2.4 Å of an acceptor from the other are found. (D) All ways of placing a glutamine residue on the peptide and any rotamer that hydrogen-bonds to it according to the criteria of (C), such that *no* pair of residues has an interaction energy of greater than 25 kcal/mol. Resultant structures were screened by eye, and if it looked plausible that three hydrogen bonds could be made to the peptidyl glutamine if the backbone were relaxed and the two indeterminate side chains were allowed sequence flexibility, then all 400 protein sequences consistent with the three determined and two indeterminate amino acids of the structure were added to the library of variant WW domains of interest.

hydrogen bonds with the glutamine residue of the ligand peptide. As reasonable-looking structures were found, all 400 protein sequences consistent with the identified 3 positions were listed by filling the two indeterminate positions with all 20 natural amino acids and added to the library of interest. As any unreasonable-looking single amino acid, pair of amino acids, or higher-order grouping was seen, all structures that contained it were removed from the set without further visual examination.

4.3 Results and Discussion

4.3.1 Individual Protein Design

Visual examination of low-energy protein sequences selected by the individual protein design algorithm indicated that variant WW domains designed for high affinity to a variant "pS-Q" ligand were unlikely to demonstrate affinity and specificity *in vitro*. Low-energy protein structures appeared to be gaining a great deal of their ligand affinity by through-solvent electrostatic interactions with the ligand phosphoserine residue. Interactions made with the glutamine were largely steric in nature, and not the type of hydrogen-bonding contacts that seem, intuitively, to be necessary for tight specific binding.

Table 4.1: **Characterization of wild-type Pin1 compared to designed WW variants.**

	Wild-type Pin1	Average of 20,000 designed sequences
Net Charge	1.00	0.85
Number of Charged Residues	1.00	3.00
Number of Heavy Atoms	29.00	23.84

In order to determine whether the lack of structures that appeared to have the capacity to bind "pS/pT-Q" ligands tightly and specifically was an artifact of our search and evaluation procedure or an effect of the physical and chemical reality for the Pin1 WW domain backbone used, we explored the relation between our low- and

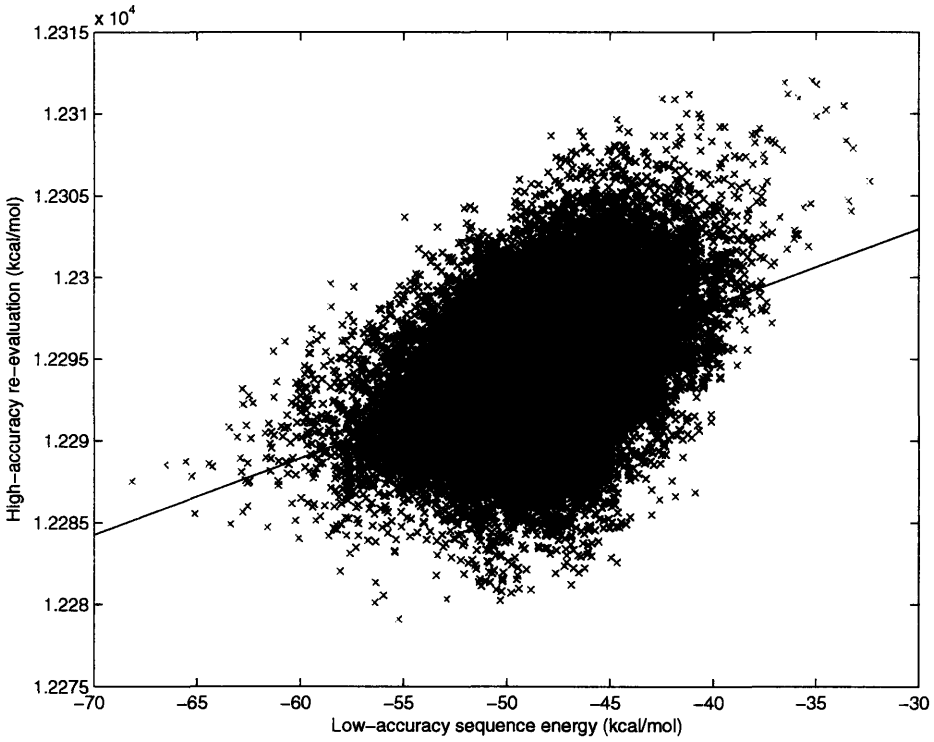
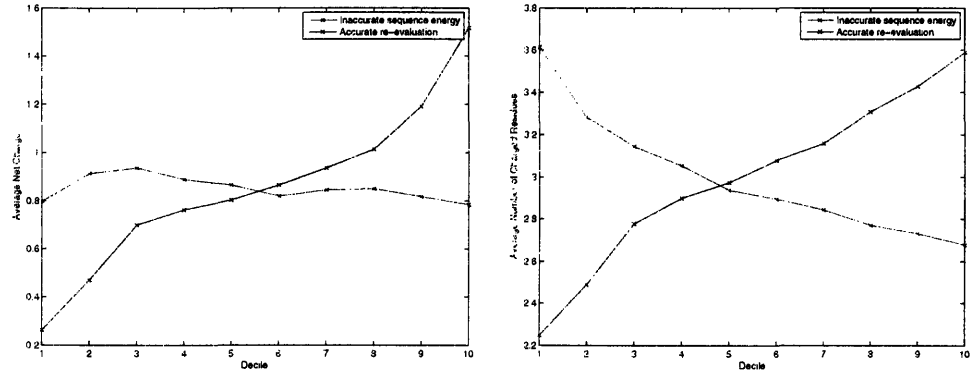


Figure 4.2: **Correlation of low- and high-accuracy energy evaluations.** The low- and high-accuracy energy evaluations of the 20,000 protein sequences with the best approximate sequence energy are plotted. Low-accuracy energy consists of internal geometry terms, $4r$ Coulombic electrostatics, and a van der Waals term. High-accuracy energy evaluation replaces the $4r$ electrostatics with a linearized Poisson–Boltzmann model, and adds a SASA-dependent hydrophobic term. The reported low- and high-accuracy energies are the best from among the ten structures with the lowest low-accuracy energy designed for each protein sequence. The plotted quantity is the difference between the folded, bound-state energy and separated unfolded model compounds. An arbitrary but constant grid-energy term is the cause of the high absolute values of high-accuracy energy evaluations. The correlation coefficient of the two evaluations is 0.41. The least-squares best-fit line is shown in black.

high-accuracy energy evaluations in this system. There is a moderate correlation (R = 0.41) between the low-accuracy energy of a protein sequence and the more accurate recalculation (see Figure 4.2), as is expected. We calculated the average net charge, number of charged residues, and number of heavy (non-hydrogen) atoms among the 5 designed WW domain amino acid positions for the 20,000 sequences analyzed with both low- and high-accuracy energy evaluation, and compared these with those of the "pS/pT-P"-binding wild-type Pin1 (see Table 4.1). We found that on average, our sequences had a lower net charge, but more charged residues than wild-type Pin1, and were smaller, with fewer heavy atoms. Though it is hard to draw conclusions from comparison to a single wild-type sequence, the fact the more charged residues are found in our designed proteins than in Pin1 may indicate a systematic tendency in our design process to design overcharged protein sequences.

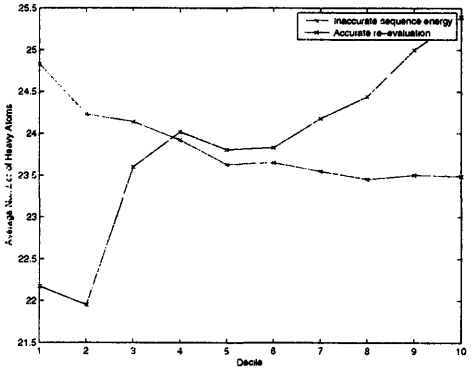
We sorted all 20000 protein sequences independently by low- and high-accuracy energy, and split each group by decile into sets of 2000. We then calculated average net charge, average number of charged residues, and average number of heavy atoms with each decile from each set (Figure 4.3). Here, the results are somewhat surprising. Despite the bulk correlation of low- and high-accuracy energies (Figure 4.2), there are some major differences in what features are highly ranked by the two functions. Average net charge (Figure 4.3A) is roughly flat across deciles in low-accuracy energy, but increases monotonically by decile when sequences are sorted on high-accuracy energies. More strikingly, average number of charged residues (Figure 4.3B) monotonically decreases by decile when sequences are sorted by low-accuracy energy, and monotonically increases when sequences are sorted by high-accuracy energy. Likewise, average number of heavy atoms per designed sequence (Figure 4.3C) tends to decrease by decile when sorted on low-accuracy energy, and rise with high-accuracy energy. Because the van der Waals component of the energy function is shared between the high- and low-accuracy evaluations, and the SASA-dependent component is very small, it seems likely that this last reversal is primarily a consequence of the fact that the positively charged amino acids are large.

On average, our pairwise low-accuracy energy evaluation highly ranks protein se-



A. Average net charge of designed residues

B. Average number of charged designed residues



C. Average number of heavy atoms among designed residues

Figure 4.3: **Characterization of designed WW variants by low- and high-accuracy energy.** Protein sequences were split into deciles by calculated low-accuracy (red) and high-accuracy (blue) energy. The (A) average net charge, (B) average number of charged residues, and (C) average number of heavy atoms within each decile are plotted.

quences that contain a large number of charged residues. On the other hand, on average, a more computationally expensive, yet more accurate, energy evaluation reranks sequences with a small number of charged residues more highly. Because an approximation to the low-accuracy function was used to identify the 20,000 protein sequences used here, it is quite possible that there are undiscovered sequences of higher low-accuracy energy that would be evaluated well by our more accurate function and that might be more likely to specifically bind "pS/pT-Q"-containing peptides. That we encountered this problem may be a function of a number of difficulties unique to the design we were performing. The presence of a highly charged anionic residue on the ligand drives the low-accuracy energy search process, which contains only a Coulombic force-field with a distance-dependent dielectric as a model of electrostatics, toward highly charged proteins; positively charged residues accumulate near the phosphoserine residue, and negatively charged residues are placed to interact favorably with the cationic residues. Moreover, the Pin1 ligand binding site redesigned here is largely solvent-exposed. This has the two effects. First, since the binding site does not constrain designed amino acid side chains very much sterically, many more protein sequences can be made to fit the binding site than would be true in a more constrained protein core. Second, the exposed site reduces the desolvation penalty paid by charged amino acids in our high-accuracy energy evaluation, permitting the overcharged interactions favored by our low-accuracy optimization function.

In addition, it is clear that complex stability and binding affinity are not necessarily good surrogates for specificity as optimization functions. Specificity is, by nature, not a pairwise decomposable function, however, making global exact optimization on binding specificity a difficult task. Indeed, in the absence of specific decoy ligands to develop specificity against, it is unclear how to proceed without using some heuristics, such as requiring that particular numbers and types of interaction are made with the ligand of interest.

4.3.2 Protein Library Design

Because a traditional single-sequence protein design proved unpromising, we attempted the design of a library of variant WW domains with potential specificity for "pS/pT-Q" peptide binding. We felt that it was important, in designing the library, to be quite liberal in allowing protein sequences with a reasonable chance at having the desired specificity into the library. On the other hand, if complete coverage of the library space by 10-fold oversampling is desired, rather than sampling, experimental library screening techniques such as yeast cell surface display [198] and phage display [199] have maximum library sizes of about 10^7 and 10^8, respectively. Moreover, any sequence added to the protein library for which there is *no* expectation of the desired specificity would act only as noise in the experimental screen. We therefore decided not to use a completely combinatorial protein library, with complete degeneracy at all five amino acid positions that we were interested in varying. Such a protein library contains only 3.2×10^6 protein sequences, although the smallest DNA library that encodes them all contains 3.3×10^7 nucleotide sequences and encodes the rarest protein sequences only once. It would, however, be quite informative to screen experimentally both the designed and a combinatorial library, and compare their relative utilities.

Table 4.2: **Designed "pS/pT-Q"-binding library.**

Residue Number					Residue Number					Residue Number					Residue Number				
23	25	27	36	38	23	25	27	36	38	23	25	27	36	38	23	25	27	36	38
ARG	XXX	ASN	ASN	XXX	ARG	XXX	ASN	ASP	XXX	ARG	XXX	ASN	GLU	XXX	ARG	XXX	ASP	ASN	XXX
ARG	XXX	ASP	ASP	XXX	ARG	XXX	ASP	GLU	XXX	ARG	XXX	GLN	XXX	GLU	ARG	XXX	GLU	ASN	XXX
ARG	XXX	XXX	ASN	LYS	LYS	XXX	ASN	ASN	XXX	LYS	XXX	ASN	ASP	XXX	LYS	XXX	ASN	GLU	XXX
LYS	XXX	ASN	XXX	LYS	LYS	XXX	ASP	ASN	XXX	LYS	XXX	ASP	ASP	XXX	LYS	XXX	ASP	GLU	XXX
LYS	XXX	ASP	XXX	LYS	LYS	XXX	GLU	ASN	XXX	LYS	XXX	GLU	XXX	LYS	LYS	XXX	XXX	ASN	LYS
LYS	XXX	XXX	ASP	LYS	LYS	XXX	XXX	GLN	LYS	LYS	XXX	XXX	GLU	LYS	XXX	ARG	ASN	ASN	XXX
XXX	ARG	ASN	ASP	XXX	XXX	ARG	ASN	GLU	XXX	XXX	ARG	ASP	ASN	XXX	XXX	ARG	ASP	ASP	XXX
XXX	ARG	ASP	GLU	XXX	XXX	ARG	GLU	ASN	XXX	XXX	GLN	ARG	XXX	ASN	XXX	GLN	ARG	XXX	ASP
XXX	GLN	ARG	XXX	GLN	XXX	GLN	ARG	XXX	GLU	XXX	GLN	ARG	XXX	HIS	XXX	GLN	ARG	XXX	THR
XXX	GLN	ASN	ASN	XXX	XXX	GLN	ASN	ASP	XXX	XXX	GLN	ASN	CYS	XXX	XXX	GLN	ASN	GLU	XXX
XXX	GLN	ASN	XXX	ASN	XXX	GLN	ASN	XXX	ASP	XXX	GLN	ASN	XXX	GLN	XXX	GLN	ASN	XXX	GLU
XXX	GLN	ASN	XXX	HIS	XXX	GLN	ASN	XXX	LYS	XXX	GLN	ASP	ASN	XXX	XXX	GLN	ASP	ASP	XXX
XXX	GLN	ASP	GLU	XXX	XXX	GLN	ASP	XXX	LYS	XXX	GLN	CYS	ASN	XXX	XXX	GLN	CYS	XXX	ASN
XXX	GLN	CYS	XXX	ASP	XXX	GLN	CYS	XXX	GLN	XXX	GLN	CYS	XXX	GLU	XXX	GLN	CYS	XXX	HIS

Continued on next page

Residue Number					Residue Number					Residue Number					Residue Number				
23	25	27	36	38	23	25	27	36	38	23	25	27	36	38	23	25	27	36	38
XXX	GLN	GLN	ASN	XXX	XXX	GLN	GLU	XXX	LYS	XXX	GLN	HIS	ASN	XXX	XXX	GLN	HIS	XXX	ASN
XXX	GLN	HIS	XXX	ASP	XXX	GLN	HIS	XXX	GLN	XXX	GLN	HIS	XXX	GLU	XXX	GLN	HIS	XXX	HIS
XXX	GLN	LYS	ASN	XXX	XXX	GLN	LYS	XXX	ASN	XXX	GLN	LYS	XXX	ASP	XXX	GLN	LYS	XXX	GLN
XXX	GLN	LYS	XXX	GLU	XXX	GLN	LYS	XXX	HIS	XXX	GLN	SER	ASN	XXX	XXX	GLN	SER	XXX	ASN
XXX	GLN	SER	XXX	ASP	XXX	GLN	SER	XXX	GLN	XXX	GLN	SER	XXX	GLU	XXX	GLN	SER	XXX	HIS
XXX	GLN	THR	ASN	XXX	XXX	GLN	THR	XXX	ASN	XXX	GLN	THR	XXX	ASP	XXX	GLN	THR	XXX	GLN
XXX	GLN	THR	XXX	GLU	XXX	GLN	THR	XXX	HIS	XXX	GLN	XXX	ASN	ASN	XXX	GLN	XXX	ASN	ASP
XXX	GLN	XXX	ASN	GLN	XXX	GLN	XXX	ASN	GLU	XXX	GLN	XXX	ASN	LYS	XXX	GLN	XXX	ASP	LYS
XXX	GLN	XXX	GLN	LYS	XXX	GLN	XXX	GLU	LYS	XXX	GLN	XXX	HIS	ASN	XXX	GLN	XXX	HIS	ASP
XXX	GLN	XXX	HIS	GLN	XXX	GLN	XXX	HIS	GLU	XXX	GLN	XXX	LYS	ASN	XXX	GLN	XXX	LYS	ASP
XXX	GLN	XXX	LYS	GLN	XXX	GLN	XXX	LYS	GLU	XXX	GLN	XXX	LYS	HIS	XXX	GLN	XXX	LYS	THR
XXX	GLU	ARG	XXX	ASN	XXX	GLU	ARG	XXX	ASP	XXX	GLU	ARG	XXX	GLN	XXX	GLU	ARG	XXX	GLU
XXX	GLU	ARG	XXX	HIS	XXX	GLU	ARG	XXX	THR	XXX	GLU	ASN	ASN	XXX	XXX	GLU	ASN	CYS	XXX
XXX	GLU	ASN	XXX	ASN	XXX	GLU	ASN	XXX	ASP	XXX	GLU	ASN	XXX	GLN	XXX	GLU	ASN	XXX	GLU
XXX	GLU	ASN	XXX	HIS	XXX	GLU	ASN	XXX	THR	XXX	GLU	CYS	ASN	XXX	XXX	GLU	CYS	GLN	XXX
XXX	GLU	CYS	LYS	XXX	XXX	GLU	CYS	XXX	ASN	XXX	GLU	CYS	XXX	ASP	XXX	GLU	CYS	XXX	GLN
XXX	GLU	CYS	XXX	GLU	XXX	GLU	CYS	XXX	HIS	XXX	GLU	CYS	XXX	THR	XXX	GLU	GLN	ASN	XXX
XXX	GLU	GLU	ASN	XXX	XXX	GLU	GLU	CYS	XXX	XXX	GLU	HIS	ASN	XXX	XXX	GLU	HIS	CYS	XXX
XXX	GLU	HIS	XXX	ASN	XXX	GLU	HIS	XXX	ASP	XXX	GLU	HIS	XXX	GLN	XXX	GLU	HIS	XXX	GLU
XXX	GLU	HIS	XXX	HIS	XXX	GLU	HIS	XXX	THR	XXX	GLU	LYS	ASN	XXX	XXX	GLU	LYS	XXX	ASN
XXX	GLU	LYS	XXX	ASP	XXX	GLU	LYS	XXX	GLN	XXX	GLU	LYS	XXX	GLU	XXX	GLU	LYS	XXX	HIS
XXX	GLU	LYS	XXX	THR	XXX	GLU	SER	ARG	XXX	XXX	GLU	SER	ASN	XXX	XXX	GLU	SER	CYS	XXX
XXX	GLU	SER	GLN	XXX	XXX	GLU	SER	LYS	XXX	XXX	GLU	SER	XXX	ASN	XXX	GLU	SER	XXX	ASP
XXX	GLU	SER	XXX	GLN	XXX	GLU	SER	XXX	GLU	XXX	GLU	SER	XXX	HIS	XXX	GLU	SER	XXX	THR
XXX	GLU	THR	ASN	XXX	XXX	GLU	THR	XXX	ASN	XXX	GLU	THR	XXX	ASP	XXX	GLU	THR	XXX	GLN
XXX	GLU	THR	XXX	GLU	XXX	GLU	THR	XXX	HIS	XXX	GLU	THR	XXX	THR	XXX	GLU	XXX	ARG	ASN
XXX	GLU	XXX	ARG	ASP	XXX	GLU	XXX	ARG	GLN	XXX	GLU	XXX	ARG	GLU	XXX	GLU	XXX	ARG	HIS
XXX	GLU	XXX	ARG	THR	XXX	GLU	XXX	ASN	ASN	XXX	GLU	XXX	ASN	ASP	XXX	GLU	XXX	ASN	GLN
XXX	GLU	XXX	ASN	GLU	XXX	GLU	XXX	ASN	THR	XXX	GLU	XXX	HIS	ASN	XXX	GLU	XXX	HIS	ASP
XXX	GLU	XXX	HIS	GLN	XXX	GLU	XXX	HIS	GLU	XXX	GLU	XXX	LYS	ASN	XXX	GLU	XXX	LYS	ASP
XXX	GLU	XXX	LYS	GLN	XXX	GLU	XXX	LYS	GLU	XXX	GLU	XXX	LYS	HIS	XXX	GLU	XXX	LYS	THR
XXX	HIS	ARG	XXX	ASN	XXX	HIS	ARG	XXX	ASP	XXX	HIS	ARG	XXX	GLN	XXX	HIS	ARG	XXX	GLU
XXX	HIS	ARG	XXX	HIS	XXX	HIS	ASN	ASN	XXX	XXX	HIS	ASN	ASP	XXX	XXX	HIS	ASN	CYS	XXX
XXX	HIS	ASN	GLU	XXX	XXX	HIS	ASN	XXX	ASN	XXX	HIS	ASN	XXX	ASP	XXX	HIS	ASN	XXX	GLN
XXX	HIS	ASN	XXX	GLU	XXX	HIS	ASN	XXX	HIS	XXX	HIS	ASN	XXX	LYS	XXX	HIS	ASP	ASN	XXX
XXX	HIS	ASP	ASP	XXX	XXX	HIS	ASP	GLU	XXX	XXX	HIS	ASP	XXX	LYS	XXX	HIS	CYS	ASN	XXX
XXX	HIS	CYS	GLN	XXX	XXX	HIS	CYS	XXX	ASN	XXX	HIS	CYS	XXX	ASP	XXX	HIS	CYS	XXX	GLN
XXX	HIS	CYS	XXX	GLU	XXX	HIS	CYS	XXX	HIS	XXX	HIS	GLN	ASN	XXX	XXX	HIS	GLU	ASN	XXX
XXX	HIS	GLU	XXX	LYS	XXX	HIS	HIS	ASN	XXX	XXX	HIS	HIS	CYS	XXX	XXX	HIS	HIS	XXX	ASN
XXX	HIS	HIS	XXX	ASP	XXX	HIS	HIS	XXX	GLN	XXX	HIS	HIS	XXX	GLU	XXX	HIS	HIS	XXX	HIS
XXX	HIS	LYS	ASN	XXX	XXX	HIS	LYS	XXX	ASN	XXX	HIS	LYS	XXX	ASP	XXX	HIS	LYS	XXX	GLN

Continued on next page

Residue Number					Residue Number					Residue Number					Residue Number				
23	25	27	36	38	23	25	27	36	38	23	25	27	36	38	23	25	27	36	38
XXX	HIS	LYS	XXX	GLU	XXX	HIS	LYS	XXX	HIS	XXX	HIS	SER	ASN	XXX	XXX	HIS	SER	GLN	XXX
XXX	HIS	SER	LYS	XXX	XXX	HIS	SER	XXX	ASN	XXX	HIS	SER	XXX	ASP	XXX	HIS	SER	XXX	GLN
XXX	HIS	SER	XXX	GLU	XXX	HIS	SER	XXX	HIS	XXX	HIS	THR	ASN	XXX	XXX	HIS	THR	XXX	ASN
XXX	HIS	THR	XXX	ASP	XXX	HIS	THR	XXX	GLN	XXX	HIS	THR	XXX	GLU	XXX	HIS	THR	XXX	HIS
XXX	HIS	XXX	ASN	GLN	XXX	HIS	XXX	ASN	GLU	XXX	HIS	XXX	ASN	LYS	XXX	HIS	XXX	ASP	LYS
XXX	HIS	XXX	GLU	LYS	XXX	HIS	XXX	LYS	ASN	XXX	HIS	XXX	LYS	ASP	XXX	HIS	XXX	LYS	GLN
XXX	HIS	XXX	LYS	GLU	XXX	HIS	XXX	LYS	HIS	XXX	LYS	ARG	XXX	ASN	XXX	LYS	ARG	XXX	ASP
XXX	LYS	ARG	XXX	GLN	XXX	LYS	ARG	XXX	GLU	XXX	LYS	ASN	ASN	XXX	XXX	LYS	ASN	ASP	XXX
XXX	LYS	ASN	GLU	XXX	XXX	LYS	ASN	XXX	LYS	XXX	LYS	ASP	ASN	XXX	XXX	LYS	ASP	ASP	XXX
XXX	LYS	ASP	GLU	XXX	XXX	LYS	ASP	XXX	LYS	XXX	LYS	GLU	ASN	XXX	XXX	LYS	GLU	XXX	LYS
XXX	LYS	XXX	ASN	LYS	XXX	LYS	XXX	ASP	LYS	XXX	LYS	XXX	GLN	LYS	XXX	LYS	XXX	GLU	LYS
XXX	LYS	XXX	LYS	GLU	XXX	XXX	ASN	ASN	ASN	XXX	XXX	ASN	ASN	ASP	XXX	XXX	ASN	ASN	GLN
XXX	XXX	ASN	ASN	LYS	XXX	XXX	ASN	ASN	THR	XXX	XXX	ASN	ASP	LYS	XXX	XXX	ASN	GLU	LYS
XXX	XXX	ASP	ASN	LYS	XXX	XXX	ASP	ASP	LYS	XXX	XXX	ASP	GLU	LYS	XXX	XXX	CYS	ASN	ASP
XXX	XXX	CYS	ASN	GLN	XXX	XXX	CYS	ASN	GLU	XXX	XXX	LYS	ASN	ASN	XXX	XXX	LYS	ASN	ASP
XXX	XXX	LYS	ASN	GLN	XXX	XXX	LYS	ASN	GLU	XXX	XXX	LYS	ASN	THR	XXX	XXX	SER	ARG	ASN
XXX	XXX	SER	ARG	ASP	XXX	XXX	SER	ARG	GLN	XXX	XXX	SER	ARG	GLU	XXX	XXX	SER	ARG	HIS
XXX	XXX	SER	ARG	THR	XXX	XXX	SER	ASN	ASN	XXX	XXX	SER	ASN	ASP	XXX	XXX	SER	ASN	GLN
XXX	XXX	SER	ASN	GLU	XXX	XXX	SER	ASN	THR	XXX	XXX	THR	ASN	GLN	XXX	XXX	THR	ASN	GLU

Table 4.2: **Designed "pS/pT-Q"-binding library.** Each element in the table represents the 400 sequences made by replacing each "XXX" with each of the 20 amino acids.

4.4 Conclusion

Here, we found 296 ways to place 3 amino acids on the Pin1 WW domain backbone such that they might possibly simultaneously hydrogen bond to a peptidyl glutamine after backbone relaxation and with sequence freedom at the other two design positions (see Table 4.2). This corresponds to 102,421 individual WW domain sequences, as some individual sequences are members of more than one of the 296 groupings. This library is more than 30-fold smaller than the combinatorial library. Although every amino acid appears at every position, there are strong interpositional correlations. Because we have included all structures in our library that might make three or more hydrogen bonds to a peptidyl glutamine, we do not expect any WW domain

sequence not in this library to bind "pS/pT-Q" ligands specifically. We have also developed a suite of software tools to attempt to express this protein library accurately as a nucleotide library for biological expression (see Chapter 5.) We have begun attempting to screen this library (see Appendix A), although significant technical hurdles remain. In combination, we are hopeful that using protein design methods to focus on relevant areas of protein sequence space and novel oligonucleotide design algorithms to express these relevant areas accurately will prove useful in the design of proteins with modified affinities, activities, and specificities.

Chapter 5

Designing Optimally Small Degenerate DNA Libraries for Accurate Expression of Protein Libraries

Abstract

The traditional mode for computational protein design has been a "design one-test one" paradigm. In many instances, it may be more productive to design a large library of related proteins and screen them experimentally. Computational protein design methods may be capable of focusing an experimental screen on a region of protein space greatly enriched for desired characteristics compared to a randomly or combinatorially mutagenized library of the same size. To address the difficulty of synthesizing a large collection of proteins spanning an irregularly shaped region of protein sequence space, we have developed a suite of tools based on the mathematical optimization method of linear integer programming to design degenerate oligonucleotide libraries that encode all desired protein sequences, and only a small number of undesired protein sequences.

This suite of tools has been applied to a large library of over 100,000 variants of the Pin1 WW domain computationally selected for the potential to bind ligand peptides containing the sequence motif "pS-Q". A number of degenerate oligonucleotide libraries are found that encode the complete set of desired variants. Interestingly, the most experimentally tractable of these is made up of just 50 single-stranded oligonucleotides, and encodes only about 11,000 additional undesired protein sequences. In comparison, the smallest traditional combinatorial library that encodes all desired protein sequences is made from only two oligonucleotides, but encodes over 3,000,000 undesired proteins, masking the desired protein library in a 30-fold excess of noise.

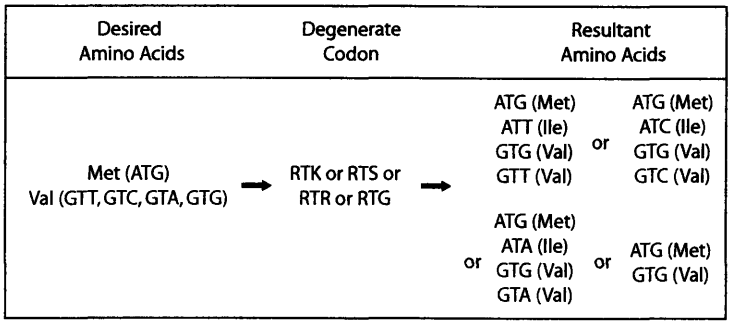
5.1 Introduction

The standard mode of operation in computational protein design has been to design, express, and test one or a small number of individual protein sequences. There has been a number of high-profile successes of protein design in this mode [122–125, 168]. Only recently have researchers begun to harness the power of computational protein design to design libraries of proteins, rather than individual sequences [178–180]. Such libraries are generally expressed in combinatorial fashion, leading to some research into methods for picking combinatorial degenerate oligonucleotide libraries (see Table 5.1 for a list of the 15 degenerate nucleotides) that best-fit the designed protein libraries [200].

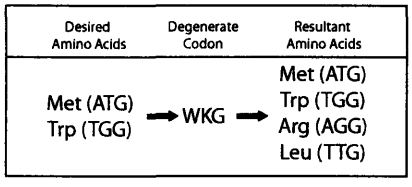
Table 5.1: **The fifteen degenerate nucleotide mixtures.**

Degenerate Nucleotide Mixture	Singular Nucleotide Composition	Degenerate Nucleotide Mixture	Singular Nucleotide Composition
A	A	C	C
T	T	G	G
R	A,G	Y	C,T
M	A,C	K	G,T
S	C,G	W	A,T
H	A,C,T	B	C,G,T
V	A,C,G	D	A,G,T
N	A,C,G,T		

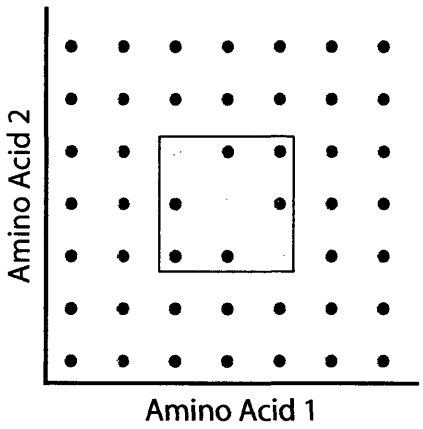
There are two major problems with attempting to express proteins from a single combinatorial degenerate oligonucleotide library (see Figure 5.1). First, the genetic code does not allow for the expression of any arbitrary list of amino acids at a single protein position with a single degenerate codon (Fig. 5.1A and B) without the addition of some undesired amino acids, or a stop codon. Second, the very nature of a combinatorial library is such that all of the combinatorially combined elements appear independently in all combinations. If the library of proteins being considered contains a great deal of interdependency among designed amino acid positions, a combinatorial oligonucleotide library that encodes all of the desired proteins will



A. Effect of The Genetic Code



B. Effect of The Genetic Code



C. Effect of Interposition Dependency

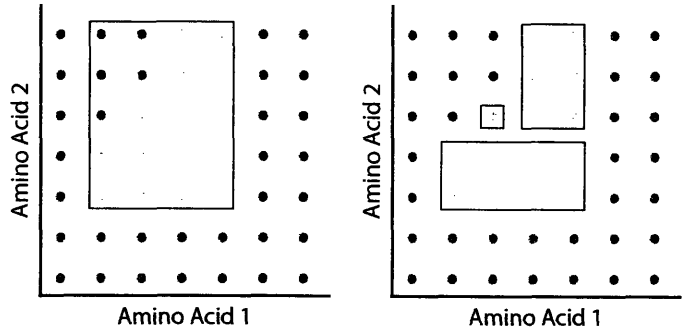
Figure 5.1: **Complications in Combinatorial Oligonucleotide Library Design.** (A) Many amino acids can be made with any of several codons. Here, in order to perfectly produce the desired amino acids valine and methionine, the degenerate codon RTG should be used, or else isoleucine will also be produced. Degenerate nucleotide definitions are as given in Table 5.1. (B) The nature of the genetic code is such that not all subsets of the amino acids can be encoded perfectly in a single position by a single degenerate codon. Here, in order to encode tryptophan (TGG) and methionine (ATG) with a single degenerate codon, the amino acids arginine (AGG) and leucine (TTG) must also be accepted. (C) The graph axes represent two amino acid positions in a protein which are being allowed to vary in sequence. When desired protein sequences (green circles) exhibit cross-positional dependency, a combinatorial library (black box) cannot contain all desired protein sequences without also containing undesired sequences (red circles). Here, the problems inherent to the genetic code are neglected for clarity.

necessarily contain undesired proteins as well (Fig 5.1C).

These undesired proteins can cause difficulty in experimental library screening at two different levels. First, experimental screening techniques are limited in the number of protein sequences that can be screened. Assuming that 10-fold redundancy is needed to screen a library, rather than sample it, phage display is useful for library sizes up to about 10^8 [199], while yeast display techniques are limited to about 10^7 unique clones [198]. These sizes correspond to protein libraries with complete degeneracy at fewer than 7 and 6 positions, respectively. Adding undesired protein sequences to a library to be screened may cause the library to grow to a size that can only be sampled, rather than exhaustively screened. Second, there is some expectation that as a library of proteins enriched for a particular function is designed, the undesired proteins left behind will be less likely than average to exhibit the desired function. Adding them into the expressed library of designed proteins adds noise to the experimental screening procedure; in effect the researcher is enlarging the haystack in which he or she will search for a needle. That is, each undesired protein sequence in the library acts as a potential false positive.

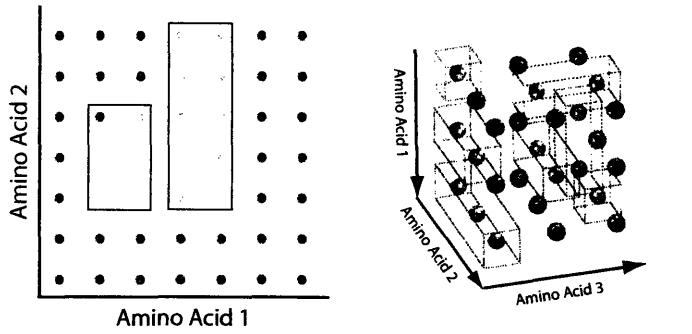
In many cases it may be desirable to create a larger number of degenerate oligonucleotides that more accurately encode the designed protein library. Here we present a method for finding the smallest possible set of degenerate oligonucleotides that encodes a set of desired proteins sequences exactly, with no extras (see Figures 5.2A and B). Moreover, we extend the method to find the smallest set of degenerate oligonucleotides encoding all desired proteins along with any predetermined allowable number of extras (see Figure 5.2C).

Properly formulated, this problem is isomorphic to a well-studied method in mathematics, the set-cover problem [201], which can be solved by linear programming. Linear programming is a means of finding the optimum of some function, subject to some constraints [201] (see Figure 5.3). Both the constraints and the objective function itself must be linear in the problem variables (Figure 5.3A). Variables can additionally be constrained to integer values [96]; this subproblem is called linear integer programming (Figure 5.3B). This framework is consistent with the problem we



A. Traditional Combinatorial Library

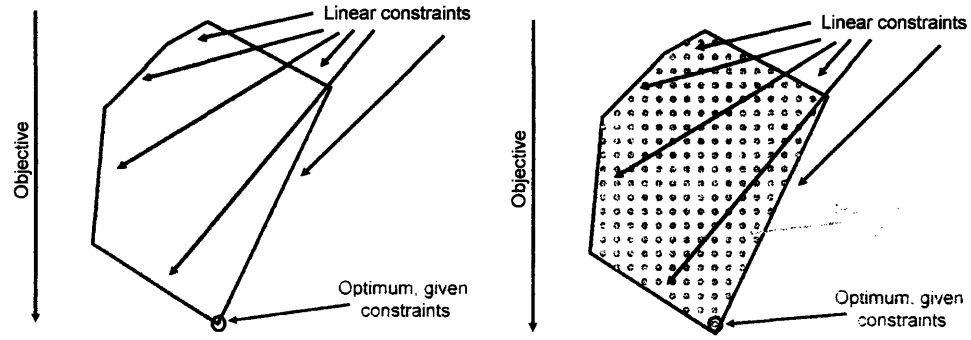
B. Exact Degenerate Library



C. Smaller Inexact Degenerate Library

D. Higher Dimensionality Exact Degenerate Library

Figure 5.2: **Oligonucleotide Library Design.** Here, graph axes represent amino acid positions in a protein which are being allowed to vary in seqence. (A-C) are two dimensional examples. (D) is a sample problem in three dimensions. Green circles represent protein sequences that the researcher is interested in screening experimentally in the shown sequence space. Red circles represent undesired sequences. Black boxes represent a single degenerate oligonucleotide that encodes the enclosed proteins. The problems inherent to the genetic code (Fig. 5.1A and B) are neglected for clarity. (A) Representation of the entire set of desired sequences with a single combinatorial degenerate oligonucleotide requires that at least five undesired protein sequences are also represented. (B) If three degenerate oligonucleotides are used, the set of desired protein sequences can be encoded, with none of the undesired sequences. Other ways of using three oligonucleotides to cover the space exactly also exist. (C) If it is acceptable to encode one undesired protein sequence, along with the fifteen desired sequences, than two degenerate oligonucleotides can be used. One other way of using two oligonucleotides to cover the same sixteen sequences exists as well. (D) Depending on the interdependency across positions, this problem can become significantly more complicated at higher dimensionality. A three-dimensional example is shown here.



A. Linear Programming **B.** Linear Integer Programming

Figure 5.3: **Linear Programming.** (A) Given a convex shape formed by a set of linear constraints on variable values, linear programming solves the problem of finding the variable values that optimize a linear objective function. (B) Linear integer programming is the same as linear programming, with some or all variables constrained to integer (and often binary) values.

---

wish to solve: minimize the number of oligonucleotides that must be made, subject to the constraints that every protein is made at least once, and no more than some limited number of undesired proteins are made. For any problem that can be formulated as an integer program, well studied algorithms and heuristics exist for finding the global optimum. There are several pre-existing software packages for identifying the solutions to linear programming problems. In this work, we use the commercial software GAMS [202, 203].

5.2 Methods

5.2.1 Protein library design

For a complete description of the design of the protein library used in this work, see Chapter 4. Briefly, a list of 102,401 variant WW domains potentially capable of binding to peptides containing the sequence motif "pS-Q" were identified by screening computationally for WW domain sequences capable of making multiple simultaneous hydrogen bonds to the glutaminyl residue of the "pS-Q" ligand. In order to increase

the computational tractability of solving the oligonucleotide design problem for this protein library, the library was broken into seven sublibraries. Each sublibrary had complete and independent degeneracy of all twenty amino acids at two positions, with interdependency at the three remaining positions. The degenerate codon *NNS*, which encodes all twenty amino acids as well as a single stop codon, was chosen and fixed at the two independently degenerate positions in each sublibrary. This was a requirement in order to make the problem solvable my a computer within the limits of the default memory allowances of the linear program solver used. The oligonucleotide library designs for the seven sublibraries were interdependently and simultaneously performed as a single linear program, such that when some undesired protein sequences are allowed into the oligonucleotide library design, the undesired sequences are distributed across the sublibraries in an optimal manner. When no undesired sequences are allowed, identical results are achieved when generating oligonucleotide libraries for the protein libraries simultaneously or independently.

5.2.2 Solution of linear integer programs

All linear integer programs were solved using the CPLEX [204] solver, in the commercial software package GAMS [202, 203].

5.2.3 Design of oligonucleotide libraries to represent protein libraries exactly

The most basic implementation of linear programming to design oligonucleotide libraries involves identifying the smallest possible set of degenerate oligonucleotides that exactly encodes a list of proteins, and no others.

Codon selection and precomputations

The space of degenerate codons is quite large – the 15 degenerate nucleotides combine to form 15^3, or 3375, degenerate codons. To simplify the formulation of the linear integer program and the computational requirements for its solution, it is extremely

useful to do some amount of precomputation to screen out codons and oligonucleotides that provably cannot be part of the final solution. First, at each mutable codon position, a list of feasible degenerate codons is generated. All codons at each position that encode any amino acid not desired at that position are removed. Also, sets of degenerate codons that coded for identical sets of amino acids are reduced, and only one is kept.

Next, oligonucleotides that could feasibly be part of the optimal set are built up by combining codons at the relevant positions, one position at a time. After each step of combination, the list of oligonucleotides is checked to ensure that each oligonucleotide will not encode any undesired proteins. This positionwise method of combining codons and checking against the list of desired proteins serves to avoid the combinatorial explosion that would come from generating all possible oligonucleotides first, and then checking against the desired protein library.

Program formulation

Once a feasible set of nonredundant oligonucleotides has been identified, a linear program can be composed to identify the smallest subset of that feasible set that encodes all desired proteins. The requirement that no undesired protein is encoded is automatically fulfilled because no individual degenerate oligonucleotide in the allowed set encodes any undesired protein. The formulation is as follows, and is exactly isomorphic to the standard set-cover problem [201]:

Minimize

$$\sum_d x_d,$$

subject to the constraints,

$$\text{for each } p, \quad \sum_d x_d \times y_{pd} \geq 1$$

$$\text{for each } d, \quad x_d \in \{0, 1\},$$

where d is the set of feasible oligonucleotides, p is the set of proteins to encode, x_d is a binary variable set to 1 if oligonucleotide d is to be part of the library, and 0 otherwise, and y_{pd} is a binary precalculated constant that is set to 1 if oligonucleotide d encodes protein p and 0 otherwise. The solution of this program ensures that the smallest possible number of oligonucleotides is used to make every protein at least once.

5.2.4 Design of oligonucleotide libraries with representation of undesired proteins allowed

Smaller oligonucleotide libraries can be generated by allowing the inclusion of undesired proteins in the expressed library (see Figure 5.2C. This requires modification of both the codon selection and the formulation of the integer program to allow the representation of protein libraries inexactly.

Codon selection and precomputations

Here, rather than keeping only codons that encode only desired amino acids, it is necessary to allow codons that encode undesired amino acids, but might ultimately yield a smaller oligonucleotide library. For every subset of amino acids desired at each position, one codon is selected that encodes those amino acids, and as few other amino acids as possible. These are combined positionwise, as above, to generate a list of feasible oligonucleotides. During the combination process, oligonucleotides

are checked to ensure that they encode a unique subset of the desired proteins. An oligonucleotide is eliminated if it encodes an identical set of desired protein to another oligonucleotide. The number of undesired proteins encoded by each oligonucleotide is also stored during this precomputation process.

Program formulation

Given this list of oligonucleotides, and knowledge of what undesired proteins they encode, and a preselected maximum number of allowable undesired protein sequences, a linear integer program can be composed that identifies the smallest subset of the oligonucleotides that encodes all desired proteins and fewer undesired proteins than a specified limit, as follows:

Minimize

$$\sum_d x_d,$$

subject to the constraints,

$$\sum_d s_d \times x_d \leq s_{max}$$

$$\text{for each } p, \quad \sum_d x_d \times y_{pd} \geq 1$$

$$\text{for each } d, \quad x_d \in \{0, 1\},$$

where d, p, x_d, and y_{pd} are the same as in the basic formulation, s_d is the precomputed number of undesired proteins encoded by oligonucleotide d, and s_{max} is the maximum allowable number of undesired proteins.

Interestingly, the nature of this program can be reversed, finding the smallest number of undesired protein sequences that can be made by a set of degenerate oligonucleotides with a given maximum size, as follows:

Minimize

$$\sum_d s_d \times x_d,$$

subject to the constraints,

$$\sum_d x_d \leq x_{max}$$

$$\text{for each } p, \quad \sum_d x_d \times y_{pd} \geq 1$$

$$\text{for each } d, \quad x_d \in \{0, 1\},$$

where x_{max} is the maximum allowable number of degenerate oligonucleotides. This formulation is useful in two instances, either when the first formulation has been run to get an minimally sized set of oligonucleotides as a way to select the best library from among all libraries of the same size, or when the researcher has a fixed limitation to the amount of money to be spent on oligonucleotide synthesis and wishes to find the best oligonucleotide library that satisfies that constraint.

Figure 5.4: **Cloning Strategy.** Forward and reverse strands are separately synthesized such that they have a small region of complementarity. Three of the five mutable positions are located on the forward strand, and two are located on the reverse (colored "X" symbols). These primers are combined and extended by PCR to form a double-stranded gene with all five mutated positions.

5.2.5 Incorporation of cloning strategy-dependent information

If it is known in advance what strategy will be used to clone the oligonucleotide library and express it as a protein library, this information can be used to design an integer program best-suited to that strategy. In the case of the sample problem discussed here, the Pin1 WW domain is small enough to be cloned fairly straightforwardly. One

coding and one anticoding strand are separately synthesized, with a small amount of overlap (see Figure 5.4). These are extended to a complete double-stranded gene by PCR, and then are ligated into an expression vector and transformed into the expressing organism. Three of the positions mutated within the protein design are located on the coding strand, and two are on the anticoding strand. Since the first three and last two mutations are encoded independently on two separate pieces of DNA, a single forward strand can be combined with multiple reverse strands, and vice versa, to encode all desired proteins. The specific pairings of forward and reverse strands returned by the solution of the linear program can be combined in individual PCR reactions, experimentally, to match the computational results. This practical consideration can be added to the formulation of an integer program of either of the types described above, either requiring an exact representation of the desired list of proteins, or allowing some prespecified number of undesired proteins. Similar modifications can be made in any case in which the experimental strategy can inform the oligonucleotide library design.

Codon selection and precomputations

While the set of allowed codons at each position is identical to that used in either of the above formulations, the precomputation of feasible degenerate oligonucleotides differs; after computing full-length double-stranded oligonucleotides, these are separated into sets of coding and anticoding single-stranded primers that contain only an appropriate subset of mutable positions.

Program formulation

Given a list of feasible coding and anticoding primers, the following integer program finds the smallest set of single-stranded primers that can be synthesized to encode all of the desired proteins, and no others. The program can be straightforwardly extended to allow any number of extra undesired proteins.

Minimize

$$\sum_f a_f + \sum_r b_r,$$

subject to the constraints,

$$\text{for each } d, \quad x_d \leq a_{f_d}$$

$$\text{for each } d, \quad x_d \leq b_{r_d}$$

$$\text{for each } p, \quad \sum_d x_d \times y_{pd} \geq 1$$

$$\text{for each } d, \quad x_d \in \{0,1\}$$

$$\text{for each } f, \quad a_f \in \{0,1\}$$

$$\text{for each } r, \quad b_r \in \{0,1\},$$

where f and r are the sets of feasible forward and reverse oligonucleotides, and a_f and b_r are binary variables indicating whether each f and r oligonucleotide is made. f_d and r_d refer to the particular members of f and r that make up a particular oligonucleotide d – these correspondences are determined prior to the composition of the linear integer program and included explicitly. All other symbols retain their meanings from above. The objective function minimizes the total number of forward and reverse primers used. The first two constraints require that a double-stranded oligonucleotide may only be used if the forward and reverse primers that it is made from are also used. The third constraint requires that each protein is made at least once, while the remaining constraints fix the x_d, a_f, and b_r variables to be binary.

5.3 Results

In order to evaluate the utility of this linear integer program method of oligonucleotide library design, the protein library designed in Chapter 4 was used as a sample case. A variety of linear programs were solved in which a wide range of numbers of undesired sequences were allowed. Moreover, programs were solved that designed double-

stranded oligonucleotides as a unit, blind to the experimental strategy by which the protein library would be expressed, as well as programs that were formulated with a knowledge of a likely expression strategy. The results are given in Figure 5.5 and Table 5.2.

Table 5.2: **Size and content of oligonucleotide libraries.**

Library Description	Number of Single-stranded Oligonucleotides	Undesired Full-length Proteins	Fraction of Proteins Desired
Traditional Libraries			
Combinatorial Library	2	3097579	0.03
Individual Protein Sequences	204842	0	1.00
Degenerate Protein Sequences	592	0	1.00
Libraries designed as double-stranded oligonucleotides			
Designed Oligo Dimers 1	138	0	1.00
Designed Oligo Dimers 2	136	400	1.00
Designed Oligo Dimers 3	132	800	0.99
Designed Oligo Dimers 4	130	1199	0.99
Designed Oligo Dimers 5	98	11597	0.90
Designed Oligo Dimers 6	46	99658	0.51
Designed Oligo Dimers 7	20	551244	0.16
Designed Oligo Dimers 8	14	1144967	0.08
Libraries designed as separate forward and reverse primers			
Designed Oligo Monomers 1	64	0	1.00
Designed Oligo Monomers 2	62	400	1.00
Designed Oligo Monomers 3	61	800	0.99
Designed Oligo Monomers 4	60	1200	0.99
Designed Oligo Monomers 5	50	10933	0.91

It is informative to compare the oligonucleotide libraries generated by this method with other, more traditional oligonucleotide libraries. Because all twenty amino acids appear at all five mutable positions of this protein library, the traditional combinatorial library approach would use two oligonucleotides (a forward and a reverse strand) to encode every amino acid at each position, for a total of 20^5 or 3,200,000 proteins (see Table 5.2, "Combinatorial Library"). Of these, 102,421 are members of the desired protein library, and 3,097,579 are not. Only 3% of the resulting proteins

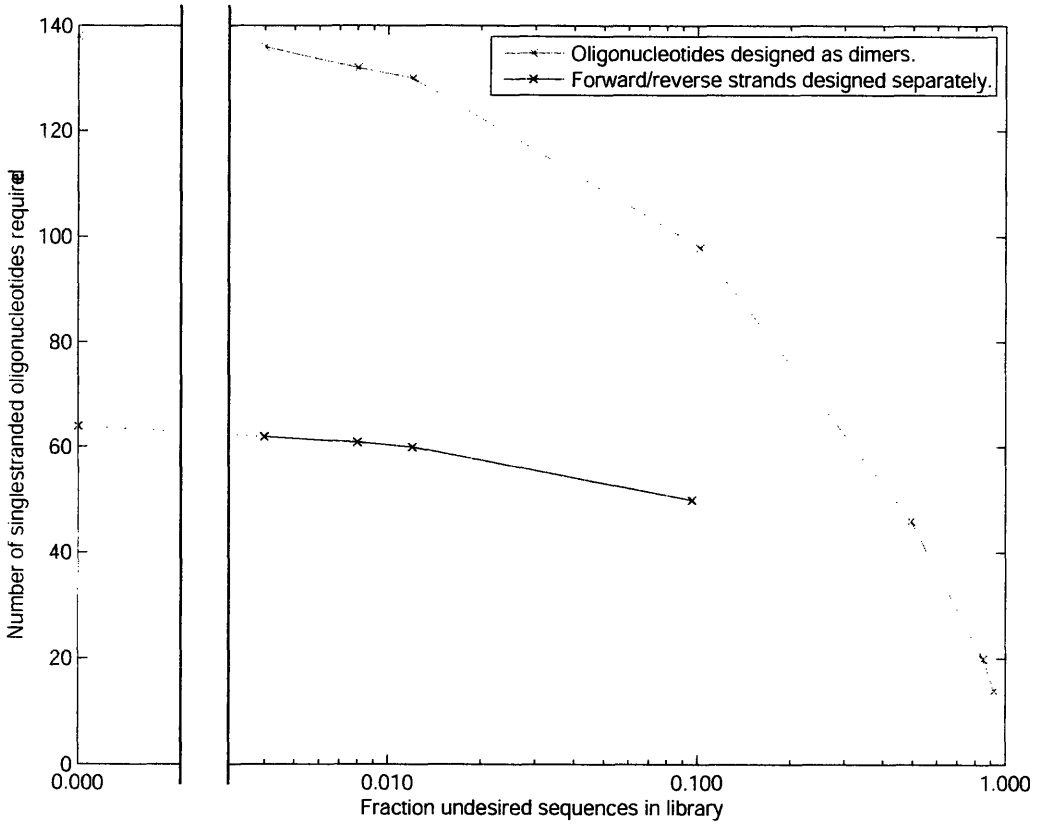


Figure 5.5: **Size and content of designed oligonucleotide libraries.** Degenerate oligonucleotide library size as a function of the number of undesired protein sequences included. The red line is composed of libraries designed as double-stranded oligonucleotides, without knowledge of experimental expression strategy. The blue line is made up of libraries designed as independent forward and reverse primers with a knowledge of how the primers would be combined experimentally to express the desired protein library. This line could not be extended further to the right without exceeding the default memory allowances of the GAMS software.

are those that were identified by protein design as being potential pSQ-binding WW domains. The rest are predicted simply to act as noise in the experimental screening process, and are in more than 30-fold excess over the desired protein sequences.

Alternatively, one might consider separately synthesizing each sequence in the protein library, or more conservatively synthesizing each of the 296 families of proteins within the library. This results in libraries of 102,421 or 296 degenerate double-stranded oligonucleotides, respectively (Table 5.2, "Individual Protein Sequences" and "Degenerate Protein Sequences"). The smaller of these is not outside the reach of some researchers, financially, although the experimental complexity of dealing with so many oligonucleotides, even if they are pooled as early as possible, is daunting. These libraries encode the entire desired protein library, and the only undesired proteins encoded are those that have a stop codon in one of the degenerate positions. These proteins are easily removed from consideration in many experimental screening procedures by requiring the presence of a C-terminal affinity tag fused to the protein of interest. The stop codons can be removed from the library at a maximum cost of 9-fold expansion of the library size; three degenerate codons are required to encode all twenty amino acids with no stop codons, and each sequence currently has two codons that encode all amino acids along with a stop codon.

In contrast, the linear integer program methodology described here is able to identify a set of 69 double-stranded oligonucleotides that encode all of the desired proteins and no others, except for those containing stop codons, as above (Table 5.2, "Designed Oligo Dimers 2"). This is less than 1/4 as many oligonucleotides as the best library described above, and certainly within experimental reach of many research groups. Moreover, the library can be made smaller still by the allowance of some number of undesired protein sequences to be encoded, along with all desired sequences. It is possible, for example, to generate a list of just 23 degenerate double-stranded oligonucleotides, or 46 monomers, that encode all desired protein sequences and a roughly equal number of undesired sequences (Table 5.2, "Designed Oligo Dimers 6"). This library contains a relatively inexpensive and easily managed number of oligonucleotides, while containing only about 1/30 as many undesired sequences as

the full combinatorial library.

Knowledge of the experimental strategy that will be used to express and screen an oligonucleotide library can be used to further reduce the size of the set of degenerate oligonucleotides needed to express a protein library. Here we have designed separate forward and reverse overlapping oligonucleotides (see Figure 5.4) that will be reused in a number of combinations to generate the full protein library. Since the forward primer contains the first three mutable positions only, and the reverse primer contains the other two, we have essentially made the two sets of mutations independently of each other, reducing the combinatorial complexity of the oligonucleotide library design. Only 64 *single-stranded* oligonucleotides are required to encode the desired set of protein sequences exactly, with no undesired sequences (Table 5.2, "Designed Oligo Monomers 1"). This is fewer than half of the monomeric oligonucleotides required to create the same protein library blind to expression strategy.

Finally, by combining the ability to allow a controlled number of undesired protein sequences into the library with separate design of forward and reverse oligonucleotides, we were able to design a set of just 50 single-stranded degenerate oligonucleotides that encode every protein sequence of interest, along with only about 10% more undesired protein sequences (Table 5.2, "Designed Oligo Monomers 5"). This library is shown in Table 5.3.

5.4 Discussion

It remains to be seen what the effect of different types and shapes of designed protein spaces is on the degenerate oligonucleotide libraries output by this method. Because any single degenerate oligonucleotide describes a many-dimensional box in protein sequence space, it seems likely that protein libraries containing densely populated and regularly shaped regions of sequence space will benefit the most from the use of libraries of multiple degenerate oligonucleotides. Indeed, the use of the methods described here on the protein library of Mena and Daugherty [200], which sparsely populates a ten-dimensional amino acid space, required almost as many oligonu-

Table 5.3: **A designed oligonucleotide library.**[a]

Forward Strands	Reverse Strands
Amino Acid Position Number	
23 25 27	36 38
NNS NNS AAK	NNS AAG
NNS NNS AGT	NNS AMC
NNS NNS RAC	NNS GAS
NNS NNS WST	NNS RAC
NNS AAG NNS	NNS SAW
NNS ARG GAA	AAC NNS
NNS CAC CAC	AAC AMC
NNS CAC WGT	AAC GAC
NNS CAK NNS	AAC SAG
NNS CAK AAC	AAG SAG
NNS CAK ARM	AAG VAC
NNS CAK VAK	ARG ACA
NNS CAK WST	CAC SAG
NNS CRM RAC	CGG VAK
NNS GAA NNS	GAA NNS
NNS GAA AMK	MAA NNS
NNS GAA VAK	MAT RAC
NNS GAA WGT	RAC NNS
NNS MAA CGG	RAC AAG
NNS MAA RAC	SAG AAG
NNS MAS GAS	TGC NNS
NNS SAG NNS	YGT NNS
NNS SAG CRC	
AAG NNS NNS	
ARG NNS NNS	
ARG NNS GAA	
ARG NNS RAC	
CGG NNS CAG	

[a] By combining one of 28 forward strands with one of 22 reverse strands in 80 different ways, all 102,421 desired protein sequences can be made, along with 10,933 undesired full-length proteins. The library also produces approximately 6% protein containing a stop codon, due to the degenerate codon "NNS", which encodes the amber ("UAG") stop codon 1/32 of the time but is allowed in this library to reduce computational complexity. Forward (coding) sequences are given for the degenerate positions on both the forward and the reverse oligonucleotides.

cleotides as proteins to exactly create the desired protein sequence space, with no extra sequences (data not shown). Further work in this direction may lead to general observations that will help in the development of methods of designing protein libraries that are particularly amenable to expression by relatively simple degenerate oligonucleotide libraries of the sort discussed here.

It is worth noting that the time-limiting step in the problem formulations, particularly as the number of variable positions increases, is the precomputation of feasible lists of oligonucleotides, which grows exponentially with the number of positions varying in protein sequence. Algorithmic improvements to this portion of the oligonucleotide library design process will be particularly useful, and may lead to an implementation fast enough to be made publicly available as a web server. Solution of integer program formulations for exact representations of protein libraries is much faster than for inexact representations for this reason.

Phrasing oligonucleotide library design as a linear integer program has particular merit, however, in that for any given formulation of the problem, a globally optimal answer can be guaranteed. The problem can be formulated such that the protein library must be encoded exactly, with no extra proteins, or with any given number of protein sequences allowed. The maximum allowable number of oligonucleotides to be purchased can be fixed, and the least noisy way of expressing the desired set of protein sequences given that constraint can be found. Moreover, knowledge of the means by which an oligonucleotide library will be expressed allows the researcher to tune a linear integer program to that strategy as a further optimization. As a result, it is possible to simultaneously tune the costs of a library-screening experiment, including monetary cost, researcher time, and the practical scientific costs of 'masking' design targets with undesired protein sequences.

Chapter 6

General Conclusions

In this work, we have developed and described a set general methods relating to the computational analysis and design of protein–protein interactions. Although we have applied the methods in almost all instances to the study of the direct interactions of phosphopeptide-binding domains with phosphorylated peptides, our methods are easily extended to other complexes.

In Chapter 2, we described the action-at-a-distance interaction, and showed that charged amino acids can interact favorably with a partner protein from outside what is traditionally considered to be the "binding interface". This is achieved by remaining solvated in the bound state, with little or no solvent-accessible surface area lost, while projecting electrostatic potential that corrects noncomplementarity at the interface. Action-at-a-distance interactions are a tempting target in tuning the affinity of a protein–protein interaction, in either direction, because their effect does not require the detailed consideration of an interfacial steric and electrostatic jigsaw puzzle, as most direct interactions do. An alternative, kinetic explanation of many of the mutations we study thermodynamically is given by Selzer *et al.* [89], who show that a number of the mutations operate primarily through an enhancement of association kinetics. Recent work has indicated that action-at-a-distance interactions are both present and designable across a wide range of protein–protein complexes [205]. Moreover, the thermodynamically enhanced action-at-a-distance mutations that are predicted by our methods and the kinetically enhanced mutations predicted by the Schreiber group's HyPare server [206] only overlap slightly, indicating that the mechanisms are largely distinct.

We moved to an analytical consideration of the requirements for phosphopeptide binding in Chapter 3. We developed a framework for building statistical models

of protein–ligand interactions, and used that framework to discover the chemical and physical properties enriched at sites of phosphoresidue contact. Despite a gross dissimilarity among the phosphate-coordinating resides of known phosphopeptide-binding domains, we quantified the propensity of phosphoresidue contact surface to be contributed by each of the twenty amino acids. We also calculated the contribution of surface curvature and electrostatic potential to phosphoresidue contact propensity. We found enrichment of a number of amino acids in sites of phosphoresidue contact. Notably, these were not only the cationic amino acids, which are common on protein surfaces in general, but also tryptophan, histidine, and tyrosine. We found that the phosphate moiety of phosphopeptides has a tendency to bind in concave protein pockets, and quantified that tendency. We also found that the negatively charged phosphate moiety bound protein surface with positive electrostatic potential, but that there was a peak in the propensity at about +8 kt/e, with more positive potential being less favorable.

After building a model of the contribution of all of these characteristics to the likelihood of phosphoresidue contact and validating this model on the known domains, we predicted the location of phosphoresidue contact on two phosphopeptide-binding domains for which the correct site was not known, the Chk1 kinase domain and the BRCA1 BRCT domain. Two predictions were made on each domain. One of the predictions on the Chk1 surface corresponded well to biochemical data [90], and the solved crystal structures of the BRCT domain of BRCA1 in complex with phosphopeptide [78, 91, 92] matched exactly one of our two predictions on that domain.

We expect the described statistical model building and site prediction methods to be generally applicable to problems involving any protein–ligand complex. The method is most appropriate to instances of convergent evolution, where the mechanism of binding is quite different among all family members. If the mechanism of action is in common among known structures, existing sequence-based and structural techniques for finding conserved motifs may be more appropriate. Care must be also be taken that appropriate propensity scales are used. A model built as ours was, to distinguish contact from non-contact surface on domains known to have contact sites,

is not sufficient to mine structure databases prospectively for novel domains, or to distinguish phosphopeptide-binding sites from other sites that have similar properties. For these more specialized tasks, more appropriate propensity scales and models must be computed. It is very likely that no single propensity scale of this sort is sufficient to mine novel phosphopeptide binding sites from the protein data bank. Instead, a first-pass model might be used to identify sites consistent with phosphopeptide binding, and a set of second models might be used to resolve true phosphopeptide-binding sites specifically from decoy sites with similar characteristics that bind other anionic ligands.

In Chapter 4, we described the design of novel phosphopeptide-binding domains. We attempt the design first of single WW domain variants with affinity and specificity for peptides containing the motif "pS/pT-Q", and then of a library of variants enriched for the same specificity relative to random variants in the same design space. Such a variant would be a useful laboratory reagent for the affinity purification of putative targets of the DNA damage kinases ATM and ATR. We evaluated the results of the individual sequence design, and found that designed structures appeared to have affinity for a "pS-Q" ligand primarily by the interaction of charged residues with the ligand phosphoserine, rather than through specific interaction with the peptidyl glutamine. There are two likely, related causes for this. The first is that complex stability relative to unfolded models was the objective set for the design. Though specificity for glutamine was not an explicit requirement of the design, it had been hoped that a designed, highly stable complex would use all available means of gaining increased stability, including tight binding between the designed domain and the peptidyl glutamine. Instead, it appeared that making long range electrostatic interactions with the peptidyl phosphoserine residue was favorable to, and to an extent precluded, making short-range interactions with glutamine. The second problem, which is related, was that the fast, inaccurate, searchable energy function used to generate designed protein sequences and structures had some systematic disagreements with a slower, more accurate energy function used to reevaluate search results. The inaccurate energy function, in particular, favored charged amino acids in a way that the accurate

energy function did not. It is possible, therefore, that if protein sequence space were searched using the slow, accurate function, the problem of favoring non-specific, long range electrostatics might be solved. Though the more accurate energy function is not pairwise-additive, and therefore not compatible with systematic, globally-optimizing search methods such as dead-end elimination and A*, it is attractive to consider using this more accurate function in a stochastic, Monte Carlo-style search starting from locations identified through the current design, to better evaluate the relative contributions of energy evaluation function and search function.

Theorizing that WW domains specific for "pS/pT-Q" would have to make hydrogen bonds to the ligand's glutamine side chain, we additionally designed a library of WW domain variants with hydrogen bonding as the primary design criterion. Using a very liberal definition of hydrogen bonding, we found all protein structures that could make 3 hydrogen bonds simultaneously to the glutamine side chain of the "pS/pT-Q" motif. We evaluated these structures by eye, and in some cases by energetic minimization, to decide whether after relaxation of the protein and peptide backbones, the hydrogen bonds could be made according to a more conservative definition. The resultant library contained 102,421 sequences. This library has not been screened experimentally, although work to evaluate the library is ongoing (see Appendix A). In developing the library, a strategy was actively pursued of including all sequences that might have the specificity of interest. This is particularly important, as the search for hydrogen-bonding structures did not account explicitly for the potential favorable effects of backbone relaxation.

Finally, in Chapter 5 we have developed methods for the design of DNA libraries capable of accurately representing protein libraries of the sort developed in Chapter 4. By using methods from the mathematical optimization field of linear integer programming, we have phrased the DNA design problem as a variant of the well-studied "set-cover" problem [201]. We have given methods for finding the smallest degenerate oligonucleotide library that encodes every protein in an arbitrary list exactly and for finding the smallest degenerate oligonucleotide library that encodes every desired protein while allowing an arbitrary, predetermined number of undesired

sequences. We also have shown how to design a degenerate oligonucleotide library of a given size that encodes all desired proteins, and the smallest possible number of undesired protein. Finally, we have shown an example of how to use knowledge about the protocol by which an oligonucleotide library will be expressed *in vivo* to inform its design. These methods were applied to the WW domain variant library described in Chapter 4. Surprisingly, we were able to design a library of just 50 degenerate single-stranded oligonucleotides that encodes all 102,421 proteins in our WW domain variant library, and only 10% as many undesired sequences.

Although the monetary cost and experimental complexity of expressing such a library are significantly greater than those of expressing a combinatorial library generated from a single oligonucleotide, there are important gains to be had from increasing the fidelity of expression of a protein library. The standard combinatorial library that encompasses all of the variant WW designs in our library encodes over 30 times as many undesired proteins as desired, and raises the probability of finding false-positive signals during experimental screening. The tools we have developed will allow researchers to find a happy medium. By using linear integer programming effectively, the monetary costs, expression complexity, and fidelity of DNA libraries that encode protein libraries can be tuned to suit the needs of the researcher. Application of this methodology to a broad variety of library design problems will help in the development of an understanding of what characteristics of a protein library make it easy or difficult to design a corresponding DNA library. This understanding may be useful in the development of protein libraries which are particularly easily expressed.

Driven by a desire to understand and design phosphodependent protein–protein interactions, we have developed a number of computational methods that we expect to have a broad utility. We have analyzed existing phosphopeptide-binding domains using a framework inspired by machine-learning algorithms, and described the properties of those domains that favor phosphopeptide binding. We have developed methods for the design of protein–protein interactions using computational biophysical methods. The action-at-a-distance interaction gives a relatively straightforward method for the tuning of protein-binding affinities. We have described some methods and

philosophy for the design of libraries of proteins with novel specificity, and explored experimental methods for screening this library. Finally, we have also developed a set of novel methods for the design of oligonucleotide libraries for the expression of protein libraries by borrowing ideas from mathematics and combinatorial optimization theory. By being willing to, and interested in, studying and developing methods in a number of fields, we have arrived at some novel solutions to problems in protein–protein complex analysis, design, and expression.

Appendix A

Testing the specificity of the Pin1 WW Domain

A.1 Introduction

The display of proteins on the yeast cell surface is a powerful technique for the detection of protein–protein interactions in a high-throughput manner [95]. In this assay, displayed protein is covalently bound to the surface of a yeast cell containing the DNA that encoded it. As with phage display [199], cells displaying proteins with the desired function can be isolated by affinity purification. However, a powerful advantage that yeast display has over phage display is the ability to assay yeast cells by fluoresence-activated cell sorting (FACS). Large libraries of 10^8 yeast cells can be rapidly sorted, either throughly searching a space of about 10^7 clones exhaustively, or sampling a larger space [198]. Cells that bind a fluorescently-labeled ligand of interest can be saved, and the identity of the protein of interest determined by sequencing following plasmid recovery. Moreover, by incubating a single clone of yeast cells in a variety of ligand concentrations, FACS can be used to determine the affinity of the protein displayed by the cells for the ligand. We have also adapted an ELISA-based assay from work by Russ *et al.* [207] to test the specificity of individual WW domains for particular phosphorylated ligands.

A.2 Materials and Methods

A.2.1 Cloning of the WW domain into pCT-CON2 in EBY100 yeast

The wild-type Pin1 WW domain (GenPept accession number NP_006212, residues 1-54) was amplified from a laboratory stock expression vector using a forward primer with an NheI endonuclease site (sequence GATGCTAGC-ATGGCGGACGAGGAGAAGCTG, NheI site underlined) and a reverse primer with a BamHI site (sequence GATGGATCCCCTGGCAGGCTCCCCCTG, BamHI site underlined). The resultant Pin1 WW gene and the yeast display vector pCT-CON2 were digested with NheI and BamHI, and ligated to generate the fusion protein [Aga2p-Xa-HA-$(G_4S)_3$-(Pin1 WW)-*c-myc*]. This vector was transformed into the EBY100 strain of yeast [95] using the method described by Schiestl and Gietz [208] for galactose-inducible cell surface display of the wild-type Pin1 WW domain.

A.2.2 Cell surface expression of the Pin1 WW domain

Colonies of transformed EBY100 were stored on selective SD-CAA (non-inducing glucose-containing growth medium) plates at 4°C. For experimental screening, a single colony was grown overnight in liquid SD-CAA at 30°C to an OD_{600} of 2.0-5.0, diluted to an OD_{600} of 1, and induced overnight at room temperature in liquid SG-CAA (inducing, galactose-containing growth medium).

A.2.3 FACS analysis of surface-displayed WW domains

0.2 OD_{600} mls of yeast cells per sample were washed in TBS and 1 mg/ml BSA, and then simultaneously incubated in varying concentrations of a labeled phospho- or nonphospho-peptide ligand, a 1:100 dilution of anti-c-Myc mouse monoclonal antibody 9E10 (Berkeley Antibody Company; Richmond, CA) to detect cell-surface display of the C-terminal c-Myc tag, or both, for times ranging from 1 to 4 hours, although we never noted any time dependence within that range. Peptide ligands

examined included biotinylated phosphphorylated and nonphosphorylated peptide libraries with the sequence Biotin-G-AHA-G-AHA-G-G-A-X-X-X-X-(phospho)T-P-X-X-X-X-A-Y-K-K-K (where X indicates any of the amino acids except cysteine, and AHA indicates aminohexanoic acid, which is used here as a linker) corresponding to the wild type Pin1 substrate motif, and FITC-labeled phosphorylated and nonphosphorylated "Pintide" with the sequence FITC-G-G-G-W-F-Y-(phospho)S-P-F-L-E-G, which has a previously reported affinity of 44 μM for Pin1 WW when phosphorylated and no detectable binding when dephosphorylated [66]. Cells were then chilled on ice, pelleted, and washed again in TBS with 1 mg/ml BSA. A secondary incubation was then performed. If a FITC-labeled peptide was used in the primary incubation, the secondary incubation was in a 1:1000 dilution of DyeMer 488/630 goat anti-mouse antibody (Molecular Probes; Eugene, OR). If a biotinylated peptide was used, secondary incubation was in a 1:100 dilution of streptavidin-conjugated phycoerythrin (Molecular Probes; Eugene, OR) and a 1:1000 dilution of FITC-conjugated goat anti-mouse antibody (Invitrogen; Carlsbad, CA). Secondary incubation was for 30 minutes, followed by a final wash in TBS and 1 mg/ml BSA. Cells were then analyzed by FACS for labeling with FITC and phycoerythrin or DyeMer 488/630.

A.2.4 ELISA analysis of Pin1 WW domain specificity

The method of Russ et al. [207] was modified as follows. DNA expressing the Pin1 WW domain (GenPept accession number NP_006212, residues 1-54) flanked by NdeI and XhoI endonuclease sites was amplified from two primers (Forward: AAGGTGCCCATATGATGGCGGACGAGGAGAAGCTGCCGCCCGGCTGGGAGAAGCGCATG ; Reverse: TGGTGGTGCTCGAGCCTGGCAGGCTCCCCCTGCCCGTTTTTGCCACCACTGCTGCTGTT, endonuclease sites underlined) and a single-stranded template (sequence CCCGGCTGGGAGAAGCGCATGAGCCGCAGCTCAGGCCGAGTGTACTACTTCAACCACATCACTAACGCCAGCCAGTGGGAGCGGCCCAGCGGCAACAGCAGCAGTGGTGGCAAA). To make a mutant WW domain with mutations only between residues 9 and 46, one would need only to replace the template. The re-

sultant DNA and a pET28-derived expression vector were digested with NdeI and XhoI and ligated to form a vector that expresses the fusion protein [GST-Pin1 WW-His_6] when induced by IPTG. This vector was transformed into and amplified in DH5α $\hat{E}$. coli cells. Plasmid was recovered using a Qiagen miniprep kit (Qiagen; Valencia, CA). Sequence-verified plasmid was transformed into the BL21 strain of *E. coli*. Cells were grown overnight, and then induced with IPTG. Cells were lysed and GST-WW was purified on glutathione-sepharose beads. The protein was eluted with 10 mM glutathione and dialyzed into TBS. Wells in a streptavidin-coated 96 well plate (Sigma-Aldrich; St. Louis, MO) were washed in TBST (TBS with 0.05% Tween-20) and incubated for 1 hour with 10 μg of a biotinylated peptide or peptide library in TBS at 4^{o}C. Wells were washed again with TBST and incubated with 1 or 100 μg of either GST, or GST-Pin1 WW in TBS for 2 hours at 4^{o}C. Wells were washed again with TBST, and incubated with a 1:5000 dilution in TBS of anti-GST antibody conjugated to horseradish peroxidase (Amersham; Piscataway, NJ) for 1 hour at 4^{o}C. Wells were washed a final time in TBST, and 100 μl of TMB solution (Sigma-Aldrich; St. Louis, MO) was added as a peroxidase ligand to each well. The reaction was stopped after 5 minutes with 50 μl of H_2SO_4, and the A_{450} of each well was read using a 96-well plate reader.

A.3 Results and Discussion

A.3.1 FACS analysis of surface-displayed WW domains

A typical FACS analysis of surface-expressed Pin1 WW domain is shown as Figures A.1 and A.2. As shown in Figure A.1, a similar percentage of cells show labeling of the C-terminal c-Myc for both the Pin1 WW domain, and D1.3, an scFv specific to hen egg lysozyme [209] (88.1% and 72.3%, respectively, cross the gate set for Pin1 C-terminal c-Myc labeling). D1.3 is used here as a negative control for phosphopeptide binding. The FITC label signal shown requires a primary incubation with anti-c-Myc antibody (data not shown). While the Pin1 WW domain appears to bind to a pT-P

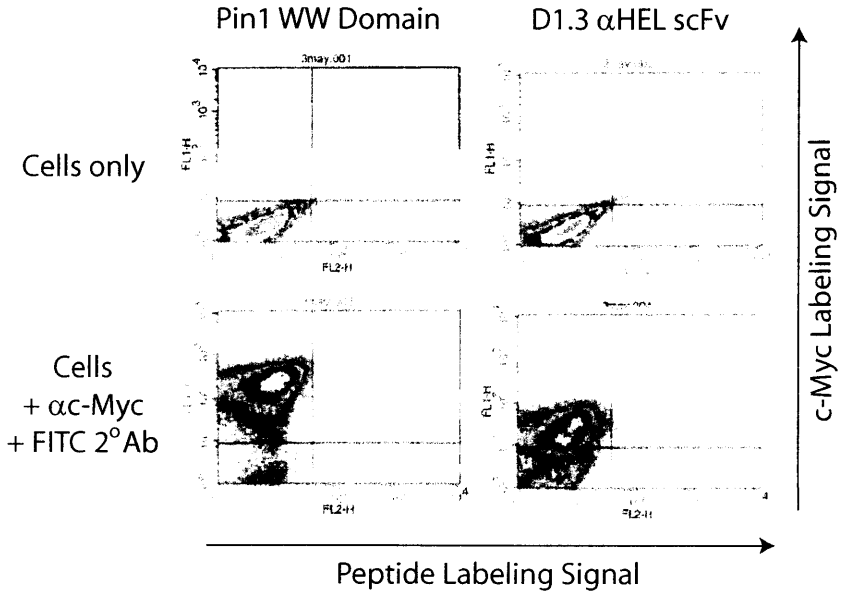


Figure A.1: **The Pin1 WW domain is cell surface expressed.** Both the Pin1 WW domain and D1.3, here used as a negative control, can be seen expressed on the yeast cell surface by the fluorescent detection of an anti-C-Myc antibody. 50,000 cells are plotted in each graph.

---

peptide library, the c-Myc-negative population of the same sample and D1.3 bind pT-P just as well, as shown in Figure A.2. No specific binding is detected.

No specific binding was likely detected for one of two reasons. First, it is possible that the WW domain is not properly folded or competent to bind phosphopeptide in the context of yeast cell surface display. This could be due to steric hindrance by the yeast display fusion construct, intracellular glycosylation by the yeast secretory pathway, or some component present or absent in the yeast extracellular environment that is different from the intracellular space where Pin1 normally functions. In the case that this is true, it is difficult to see how to proceed with a yeast display methodology, although altering induction conditions might prompt more accurate folding of the domain. It is reported, however, that the Pin1 WW domain is stably folded on its own in solution [210], so any misfolding or binding-site obstruction in this system is likely to be system-induced, and difficult to escape.

Second, and perhaps more likely, it is possible that the dissociation kinetics of the

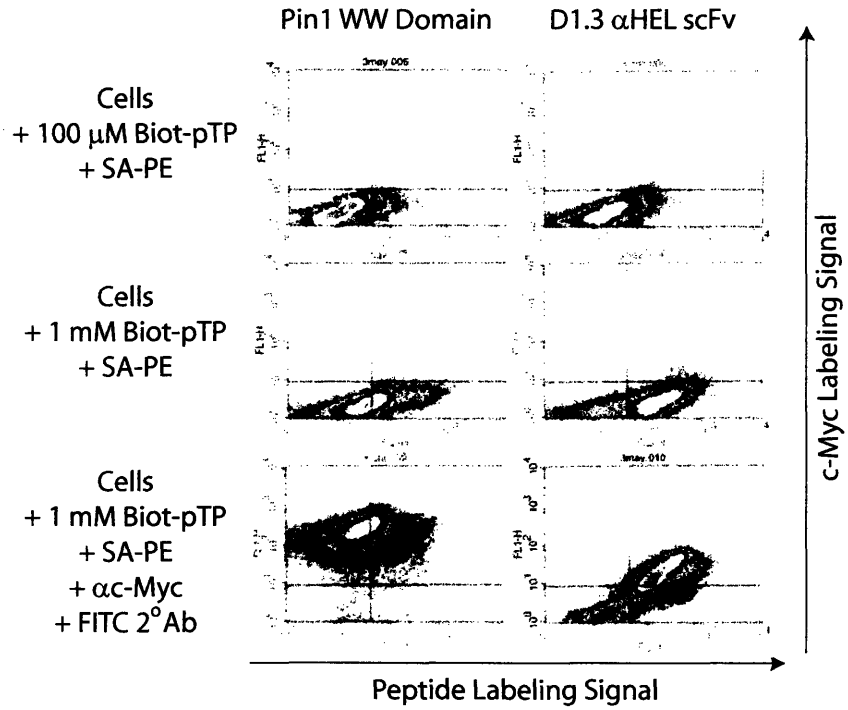


Figure A.2: **The surface expressed Pin1 WW domain does not bind "pT-P" peptide library specifically.** While biotinylated "pT-P" peptide library sticks to cells surface expressing the Pin1 WW domain, it sticks as well to the expression-negative subpopulation of the same sample, and to D1.3. No specific phosphopeptide binding is detected. 50,000 cells are plotted in each graph.

Pin1 WW domain are too fast for examination by this protocol. The reported affinity of the Pin1 WW domain for phosphorylated Pintide is quite weak, just 44 μM [66]. Protein-protein association rates are generally in the vicinity of 10^5-10^6 $M^{-1}s^{-1}$, and may be faster here given the small protein and peptide ligand involved. This gives an estimated dissociation rate of 4.4-44 s^{-1}, corresponding to a complex half-life of about 0.02-0.16 s in dilute solution. It is possible, therefore that it will be a necessity in the future to study WW domain/peptide interactions only under equilibrium conditions, or in other ways where the possibility of an extremely high off-rate is not a hindrance. It is also possible that by utilizing avidity effects, either by surface expressing tandem WW domains, or by creating large clusters of ligand peptides in close proximity, as with MAP peptides [211].

A.3.2 ELISA analysis of Pin1 WW domain specificity

While not well suited to the screening of a WW domain variant library, we have also recently tested an ELISA assay for the elucidation of the specificity of individual WW domains, roughly as described in [207] (see Figure A.3). We found that more active horseradish peroxidase conjugated to an anti-GST antibody was found in wells where GST-fused Pin1 WW domain had been incubated with peptides and peptide libraries containing a "pS/T-P" motif than with "pS/T-Q" or "S/T-P" motifs. This indicates that Pin1 WW domain is specific for "pS/T-P" relative to these other motifs, as expected [54, 66]. This assay is sufficient to test the specificity of any individual WW domains designed in the future for "pS/T-P" relative to "pS/T-Q".

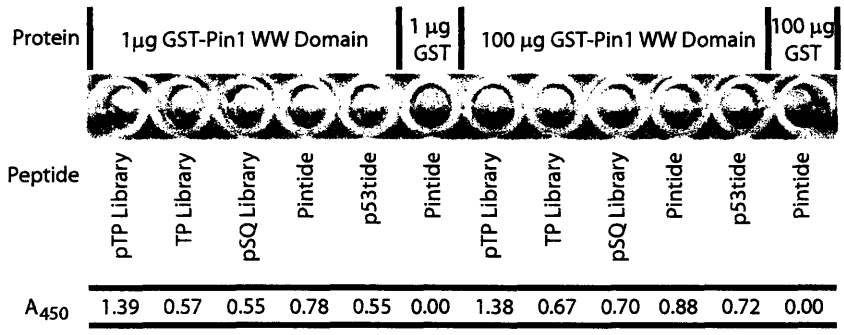


Figure A.3: **A GST-Pin1 WW domain fusion is specific for "pS/T-P" peptides and peptide libraries.** The Pin1 WW domain shows specificity to "pT-P" library, relative to "pS-Q" library and "T-P" library, and to Pintide relative to a "pS-Q" peptide derived from a site on the protein p53. The A_{450} of GST alone is subtracted from all samples at the same protein concentration. The sequences of the peptides used are:
pTP Library: Biotin-G-AHA-G-AHA-G-G-A-X-X-X-X-pT-P-X-X-X-X-A-Y-K-K-K
TP Library: Biotin-G-AHA-G-AHA-G-G-A-X-X-X-X-T-P-X-X-X-X-A-Y-K-K-K
pSQ Library: Biotin-G-AHA-G-AHA-G-G-A-X-X-X-X-pS-Q-X-X-X-X-A-Y-K-K-K
Pintide: Biotin-G-AHA-G-AHA-G-G-W-F-Y-pS-P-F-L-E-A-Y-K-K-K
p53tide: Biotin-G-AHA-G-AHA-G-G-E-P-P-L-pS-Q-E-T-F-A-Y-K-K-K

Bibliography

[1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[2] P.T.W. Cohen. Protein phosphatase 1 - targeted in many directions. *J. Cell Sci.*, 115:241–256, 2002.

[3] G.M. Rubin, M.D. Yandell, J.R. Wortman, G.L. Gabor Miklos, C.R. Nelson, I.K. Hariharan, M.E. Fortini, P.W. Li, R. Apweiler, W. Fleischmann, J.M. Cherry, S. Henikoff, M.P. Skupski, S. Misra, M. Ashburner, E. Birney, M.S. Boguski, T. Brody, P. Brokstein, S.E. Celniker, S.A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R.F. Galle, W.M. Gelbart, R.A. George, L.S.B. Goldstein, F. Gong, P. Guan, N.L. Harris, B.A. Hay, R.A. Hoskins, J. Li, Z. Li, R.O. Hynes, S.J.M. Jones, P.M. Kuehl, B. Lemaitre, J.T. Littleton, D.K. Morrison, C. Mungall, P.H. O'Farrell, O.K. Pickeral, C. Shue, L.B. Vosshall, J. Zhang, Q. Zhao, X.H. Zheng, F. Zhong, W. Zhong, R. Gibbs, J.C. Venter, M.D. Adams, and S. Lewis. Comparative genomics of the eukaryotes. *Science*, 287:2204–2215, 2000.

[4] M. Huse and J. Kuriyan. The conformational plasticity of protein kinases. *Cell*, 109:275–282, 2002.

[5] S. R. Hubbard. Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J.*, 16:5573–5581, 1997.

[6] B. Canagarajah, A. Khokhlatchev, M.H. Cobb, and E.J. Goldsmith. Activation

mechanism of the MAP kinase ERK2 by dual phosphorylation. *Cell*, 90:859–869, 1997.

[7] L.N. Johnson and R.J. Lewis. Structural basis for control by phosphorylation. *Chem. Rev.*, 101:2209–2242, 2001.

[8] P.D. Jeffrey, A.A. Russo, K. Polyak, E. Gibbs, J. Hurwitz, J. Massagué, and N.P. Pavletich. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature*, 376:313–320, 1995.

[9] I. Moarefi, M. LaFevre-Bernt, F. Sicheri, M. Huse, C.-H. Lee, J. Kuriyan, and W.T. Miller. Activation of the Src-family kinase Hck by SH3 domain displacement. *Nature*, 385:650–653, 1997.

[10] S. Gonfloni, A. Weijland, J. Kretzschmar, and G. Superti-Furga. Crosstalk between the catalytic and regulatory domains allows bidirectional regulation of Src. *Nat. Struct. Biol.*, 7:281–286, 2000.

[11] D.R. Knighton, J. Zheng, L.F. TenEyck, V.A. Ashford, N.-H. Xuong, S.S. Taylor, and J.M. Sowadski. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science*, 253:407–414, 1991.

[12] S.-H. Hu, M.W. Parker, J.Y. Lei, M.C.J. Wilce, G.M. Benian, and B.E. Kemp. Insights into autoregulation from the crystal structure of twitchin kinase. *Nature*, 369:581–584, 1994.

[13] J. Goldberg, A.C. Nairn, and J. Kuriyan. Structural basis for the autoinhibition of calcium/calmodulin-dependent protein kinase I. *Cell*, 84:875–887, 1996.

[14] O. Mayans, P.F. van der Ven, M. Wilm, A. Mues, P. Young, D.O. Furst, M. Wilmanns, and M. Gautel. Structural basis for activation of the titin kinase domain during myofibrillogenesis. *Nature*, 396:863–869, 1998.

[15] M. Lei, W¿ Lu, W. Meng, M.C. Parrini, M.J. Eck, B.J. Mayer, and S.C. Harrison. Structure of pak1 in an autoinhibited conformation reveals a multistage activation switch. *Cell*, 102:387–397, 2000.

[16] J. Zimmermann, E. Buchdunger, H. Mett, T. Meyer, and N.B. Lydon. Potent and selective inhibitors of the Abl-kinase: phenylamino-pyrimidine (PAP) derivatives. *Bioorg. Med. Chem. Lett.*, 7:187–192, 1997.

[17] B.J. Druker, M. Talpaz, D.J. Resta, B. Peng, E. Buchdunger, J.M. Ford, N.B. Lydon, H. Kantarjian, R. Capdeville, S. Ohno-Jones, and C.L. Sawyers. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.*, 344:1031–1037, 2001.

[18] P. J. Kennelly and E. G. Krebs. Consensus sequences as substrate specificity determinants for protein kinases and protein phosphatases. *J. Biol. Chem.*, 266:15555–15558, 1991.

[19] P. Cohen. The search for physiological substrates of MAP and SAP kinases in mammalian cells. *Trends Cell Biol.*, 7:353–361, 1997.

[20] K. Shah, Y. Liu, C. Deirmengian, and K. M. Shokat. Engineering unnatural nucleotide specificity for rous sarcoma virus tyrosine kinase to uniquely label its direct substrates. *Proc. Natl. Acad. Sci. U.S.A.*, 94:3565–3570, 1994.

[21] Y. Liu, K. Shah, F. Yang, L. Witucki, and K. M. Shokat. Engineering Src family protein kinases with unnatural nucleotide specificity. *Curr. Biol.*, 5:91–102, 1998.

[22] A. Y. Ting, K. Witte, K. Shah, B. Kraybill, K. M. Shokat, and P. G. Schultz. Phage-display evolution of tyrosine kinases with altered nucleotide specificity. *Biopolymers (Pept. Sci.)*, 60:220–228, 2001.

[23] K. Shah and K. M. Shokat. A chemical genetic screen for direct v-Src substrates reveals ordered assembly of a retrograde signaling pathway. *Curr. Biol.*, 9:35–47, 2002.

[24] G. Manning, D.B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298:1912–1934, 2002.

[25] J. Rush, A. Moritz, K.A. Lee, A. Guo, V.L. Goss, E.J. Spek, H. Zhang, X.-M. Zha, R.D. Polakiewicz, and M.J. Comb. Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.*, 23:94–101, 2004.

[26] Y. Zhang, A. Wolf-Yadlin, P.L. Ross, D.J. Pappin, J. Rush, and F.M. Lauffenburger, D.A. White. Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. *Mol. Cel. Proteomics*, 4:1240–1250, 2005.

[27] M. de Graauw, P. Hensbergen, and B. van de Water. Phospho-proteomic analysis of cellular signaling. *Electrophoresis*, 27:2676–2686, 2006.

[28] Z. Songyang, S. Blechner, N. Hoagland, M. F. Hoekstra, H. Piwnica-Worms, and L. C. Cantley. Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr. Biol.*, 4:973–982, 1994.

[29] Z. Songyang and L.C. Cantley. The use of peptide library for the determination of kinase peptide substrates. *Methods Mol. Biol.*, 87:87–98, 1998.

[30] J.E. Hutti, E.T. Jarrell, J.D. Chang, D.W. Abbod, P. Storz, A. Toker, L.C. Cantley, and B.E. Turk. A rapid method for determining protein kinase phosphorylation specificity. *Nat. Methods*, 1:27–29, 2004.

[31] R. Gast, J. Glökler, M. Höxter, M. Kieß, R. Frank, and W. Tegge. Method for determining protein kinase substrate specificities by the phosphorylation of peptide libraries on beads, phosphate-specific staining, automated sorting, and sequencing. *Anal. Biochem.*, 276:227–241, 1999.

[32] T. P. Cujec, P. F. Medeiros, P. Hammond, C. Rise, and B. L. Kreider. Selection of v-Abl tyrosine kinase substrate sequences from randomized peptide and cellular proteomic libraries using mRNA display. *Chem. & Biol.*, 9:253–264, 2002.

[33] M. B. Yaffe, G. G. Leparc, J. Lai, T. Obata, S. Volinia, and L. C. Cantley. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol.*, 19:348–353, 2001.

[34] J.C. Obenauer, L.C. Cantley, and M.B. Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, 31:3635–3641, 2003.

[35] L.N. Johnson. Glycogen phosphorylase: control by phosphorylation and allosteric affectors. *FASEB J.*, 6:2274–2282, 1992.

[36] S.R. Hubbard, L. Wei, L. Ellis, and W.A. Hendrickson. Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature*, 372:746–754, 1994.

[37] B. Derijard, J. Raingeaud, T. Barrett, I.H. Wu, J. Han, R.J. Ulevitch, and R.J. Davis. Independent human MAP-kinase signal transduction pathways defined by MEK and MKK isoforms. *Science*, 267:682–685, 1995.

[38] M. B. Yaffe and L. C. Cantley. Grabbing phosphoproteins. *Nature*, 402:30–31, 1999.

[39] M.-M. Zhou. Phosphothreonine recognition comes into focus. *Nat. Struct. Biol.*, 7:1085–1087, 2000.

[40] M.B. Yaffe and A.E. Elia. Phosphoserine/threonine-binding domains. *Curr. Opin. Cell Biol.*, 13:131–138, 2001.

[41] M. B. Yaffe and S. J. Smerdon. PhosphoSerine/Threonine binding domains: You can't pSERious? *Structure*, 9:R33–R38, 2001.

[42] M. B. Yaffe. Phosphotyrosine-binding domains in signal transduction. *Nat. Rev.: Mol. Cell Biol.*, 3:177–185, 2002.

[43] M. Matsuda, B.J. Mayer, Y. Fukui, and H. Hanafusa. Binding of transforming protein, $p47^{gag-crk}$, to a broad range of phosphotyrosine-containing peptides. *Science*, 248:1537–1539, 1990.

[44] D. Anderson, C.A. Koch, L. Grey, C. Ellis, M.F. Moran, and T. Pawson. Binding of SH2 domains of phospholipase $C_\gamma 1$, GAP and Src to activated growth factor receptors. *Science*, 250:979–982, 1990.

[45] M.F. Moran, C.A. Koch, D. Anderson, C. Ellis, L. England, G.S. Martin, and T. Pawson. Src homology region-2 domains direct protein-protein interactions in signal transduction. *Proc. Natl. Acad. Sci. USA*, 87:8622–8626, 1990.

[46] M. Matsuda, B.J. Mayer, and H. Hanafusa. Identification of domains of the *v*-crk oncogene product sufficient for association with phosphotyrosine-containing proteins. *Mol. Cell. Biol.*, 11:1607–1613, 1991.

[47] B.J. Mayer, P.K. Jackson, and D. Baltimore. The noncatalytic src homology region 2 segment of abl tyrosine kinase binds to tyrosine-phosphorylated cellular proteins with high affinity. *Proc. Natl. Acad. Sci. USA*, 88:627–631, 1991.

[48] J.M. Bradshaw, V. Mitaxov, and G. Waksman. Investigation of phosphotyrosine recognition by the SH2 domain of the Src kinase. *J. Mol. Biol.*, 293:971–985, 1999.

[49] P. Blaikie, D. Immanuel, J. Wu, N.X. Lu, V. Yajnik, and B. Margolis. A region in Shc distinct from the SH2 domain can bind tyrosine-phosphorylated growth factor receptors. *J. Biol. Chem.*, 269:32031–32034, 1994.

[50] W.M. Kavanaugh and L.T. Williams. An alternative to SH2 domains for binding tyrosine-phosphorylated proteins. *Science*, 266:1862–1865, 1994.

[51] A.J. Muslin, J.W. Tanner, P.M. Allen, and A.S. Shaw. Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine. *Cell*, 84:889–897, 1996.

[52] M. B. Yaffe, K. Rittinger, S. Volinia, P. R. Caron, A. Aitken, H. Leffers, S. J. Gamblin, S. J. Smerdon, and L. C. Cantley. The structural basis of 14-3-3:phosphopeptide binding specificity. *Cell*, 91:961–971, 1997.

[53] B. Xiao, S.J. Smerdon, D.H. Jones, G.G. Dodson, Y. Soneki, A. Aitken, and S.J. Gamblin. Structure of a 14-3-3 protein and implications for coordination of multiple signaling pathways. *Nature*, 376:188–191, 1995.

[54] P.J. Lu, X.Z. Zhou, M. Shen, and K.P. Lu. A function of WW domains as phosphoserine- or phosphothreonine-binding modules. *Science*, 283:1325–1328, 1999.

[55] M.B. Yaffe, M. Schutkowski, M.H. Shen, X.Z. Zhou, P.T. Stukenberg, J.U. Rahfeld, J. Xu, J. Kuang, M.W. Kirschner, G. Fischer, Lu K.P., and L.C. Cantley. Sequence-specific and phosphorylation-dependent proline isomerization: A potential mitotic regulatory mechanism. *Science*, 278:1957–1960, 1997.

[56] D. Durocher, J. Henckel, A.R. Fersht, and S.P. Jackson. The FHA domain is a modular phosphopeptide recognition motif. *Mol. Cell*, 4:387–394, 1999.

[57] A.E.H. Elia, L.C. Cantley, and M.B. Yaffe. Proteomic screen finds pSer/pThr-binding domain localizing Plk1 to mitotic substrates. *Science*, 299:1228–1231, 2003.

[58] I.A. Manke, D.M. Lowery, A. Nguyen, and M.B. Yaffe. BRCT repeats as phosphopeptide-binding modules involved in protein targeting. *Science*, 302:636–639, 2003.

[59] Z. Songyang, S.E. Shoelson, M. Chaudhuri, G. Gish, T. Pawson, W.G. Haser, F. King, T. Roberts, S. Ratnofsky, R.J. Lechleider, B.G. Neel, B. Birge, J.E. Fajardo, M.M. Chou, H. Hanafusa, B. Schaffhausen, and L.C. Cantley. SH2 domains recognize specific phosphopeptide sequences. *Cell*, 72:767–778, 1993.

[60] Z. Songyang, B. Margolis, M. Chaudhuri, S.E. Shoelson, and L.C. Cantley. The

phosphotyrosine interaction domain of SHC recognizes tyrosine-phosphorylated NPXY motif. *J. Biol. Chem*, 270, 1995.

[61] D. Durocher, I. A. Taylor, D. Sarbassova, L. F. Haire, S. L. Westcott, S. P. Jackson, S. J. Smerdon, and M. B. Yaffe. The molecular basis of FHA domain:phosphopeptide binding specificity and implications for phospho-dependent signaling mechanisms. *Mol. Cell*, 6:1169–1182, 2000.

[62] J.T. Winston, P. Strack, P. Beer-Romero, C.Y. Chu, S.J. Elledge, and J.W. Harper. The $SCF^{\beta-TRCP}$-ubiquitin ligase complex associates specifically with phosphorylated destruction motifs in IκBα and β-catenin and stimulates IκBα ubiquitination *in vitro*. *Genes Dev.*, 13:270–283, 1999.

[63] J.W. Wu, M. Hu, J.J. Chai, J. Seoane, M. Huse, C. Li, D.J. Rigotti, S. Kyin, T.W. Muir, R. Fairman, J. Massague, and Y. Shi. Crystal structure of a phosphorylated Smad2: Recognition of phosphoserine by the MH2 domain and insights on Smad function in TGF-beta signaling. *Mol. Cell*, 8:1277–1289, 2001.

[64] A.E.H. Elia, P. Rellos, L.F. Haire, J.W. Chao, F.J. Ivins, K. Hoepker, D. Mohammad, L.C. Cantley, S.J. Smerdon, and M.B. Yaffe. The molecular basis for phosphodependent substrate targeting and regulation of Plks by the Polo-box domain. *Cell*, 115:83–95, 2003.

[65] K. Rittinger, J. Budman, J.A. Xu, S. Volinia, L.C. Cantley, S.J. Smerdon, S.J. Gamblin, and M.B. Yaffe. Structural analysis of 14-3-3 phosphopeptide complexes identifies a dual role for the nuclear export signal of 14-3-3 in ligand binding. *Mol. Cell*, 4:153–166, 1999.

[66] M. A. Verdecia, M. E. Bowman, K. P. Lu, T. Hunter, and J. P. Noel. Structural basis for phosphoserine-proline recognition by group IV WW domains. *Nat. Struct. Biol.*, 7:639–643, 2000.

[67] M.J. Eck, S.E. Shoelson, and S.C. Harrison. Recognition of a high-affinity

phosphotyrosyl peptide by the Src homology-2 domain of p56(Lck). *Nature*, 362:87–91, 1993.

[68] C.Y. Peng, P.R. Graves, R.S. Thoma, Z. Wu, A.S. Shaw, and H. Piwnica-Worms. Mitotic and G2 checkpoint control: regulation of 14-3-3 protein binding by phosphorylation of Cdc25C on serine-216. *Science*, 277:1501–1505, 1997.

[69] Y. Zeng and H. Piwnica-Worms. DNA damage and replication checkpoints in fission yeast require nuclear exclusion of the Cdc25 phosphatase via 14-3-3 binding. *Mol. Cell Biol.*, 19:7410–7419, 1999.

[70] J. Yang, K. Winkler, M. Yoshida, and S. Kornbluth. Maintenance of G2 arrest in the *Xenopus* oocyte: a role for 14-3-3 mediated inhibition of Cdc25 nuclear import. *EMBO J.*, 18:2174–2183, 1999.

[71] A. Kumagai and W.G. Dunphy. Binding of 14-3-3 proteins and nuclear export control the intracellular localization of the mitotic inducer Cdc25. *Genes Dev.*, 13:1067–1072, 1999.

[72] A. Lopez-Girona, B. Furnari, O. Mondesert, and P. Russell. Nuclear localization of Cdc25 is regulated by DNA damage and a 14-3-3 protein. *Nature*, 397:172–175, 1999.

[73] D.P. Morris, H.P. Phatnani, and A.L. Greenleaf. Phospho-carboxyl-terminal domain binding and the role of a prolyl isomerase in pre-mRNA 3'-end formation. *J. Biol. Chem.*, 274:31583–31587, 1999.

[74] F.A. Barr, H.H.W. Silljé, and E.A. Nigg. Polo-like kinases and the orchestration of cell division. *Nat. Rev. Mol. Cell Biol.*, 5:429–441, 2004.

[75] D.M. Lowery, D. Lim, and M.B. Yaffe. Structure and function of Polo-like kinases. *Oncogene*, 24, 2005.

[76] X.C. Yu, C.C.S. Chini, M. He, G. Mer, and J.J. Chen. The BRCT domain is a phospho-protein binding domain. *Science*, 302:639–642, 2003.

[77] R. Scully and D.M. Livingston. In search of the tumour-suppressor functions of BRCA1 and BRCA2. *Nature*, 408:429–432, 2000.

[78] J.A. Clapperton, I.A. Manke, D.M. Lowery, T. Ho, L.F. Haire, M.B. Yaffe, and S.J. Smerdon. Structure and mechanism of BRCA1 BRCT domain recognition of phosphorylated BACH1 with implications for cancer. *Nat. Struct. Mol. Biol.*, 11:512–518, 2004.

[79] M. Pozuelo Rubio, K.M. Geraghty, B.H.C. Wong, N.T. Wood, D.G. Campbell, N. Morrice, and C. Mackintosh. 14-3-3-affinity purification of over 200 human phosphoproteins reveals new links to regulation of cellular metabolism, proliferation and trafficking. *Biochem. J.*, 379:395–408, 2004.

[80] M.B. Yaffe and L.C. Cantley. Mapping specificity determinants for protein–protein association using protein fusions and random peptide libraries. *Meth. Enzymol.*, 328:157–170, 2000.

[81] D.W. Powell, M.J. Rane, B.A. Joughin, R. Kalmukova, J.H. Jong, B. Tidor, W.L. Dean, W.M. Pierce, J.B. Klein, M.B. Yaffe, and K.R. McLeish. Proteomic identification of 14-3-3 zeta as a mitogen-activated protein kinase-activated protein kinase 2 substrate: Role in dimer formation and ligand binding. *Mol. Cell. Biol.*, 23:5376–5387, 2003.

[82] D.V. Bulavin, Y. Higashimoto, Z.N. Bemidenko, S. Meek, P. Graves, C. Phillips, H. Zhao, S.A. Moody, E. Appella, H. Piwnica-Worms, and A.J. Fornace. Dual phosphorylation controls Cdc25 phosphatases and mitotic entry. *Nat. Cell Biol.*, 5:545–551, 2003.

[83] A.M. Brownawell, G.J.P.L. Kops, I.G. Macara, and B.M.T. Burgering. Inhibition of nuclear import by protein kinase B (Akt) regulates the subcellular distribution and activity of the forkhead transcription factor AFX. *Mol. Cell. Biol.*, 21:3534–3546, 2001.

[84] A. Brunet, F. Kanai, J. Stehn, J. Xu, D. Sarbassova, J.V. Frangioni, S.N. Dalal, J.A. DeCaprio, M.E. Greenberg, and M.B. Yaffe. 14-3-3 transits to the nucleus and participates in dynamic nucleoplasmic transport. *J. Cell Biol.*, 156:817–828, 2002.

[85] J.B. Aggen, A.C. Nairn, and R. Chamberlin. Regulation of protein phosphatase-1. *Chem. Biol.*, 7:R13–R23, 2000.

[86] M. Bollen. Combinatorial control of protein phosphatase-1. *Trends Biochem. Sci.*, 26:426–431, 2001.

[87] A. Alonso, J. Sasin, N. Bottini, I. Friedberg, I. Friedberg, A. Osterman, A. Godzik, T. Hunter, J. Dixon, and T. Mustelin. Protein tyrosine phosphatases in the human genome. *Cell*, 117:600–711, 2004.

[88] T.E. Creighton. *Proteins: Structures and Molecular Properties.* W. H. Freeman, New York, 1992.

[89] T. Selzer, S. Albeck, and G. Schreiber. Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.*, 7:537–541, 2000.

[90] S.Y. Jeong, A. Kumagai, J. Lee, and W.G. Dunphy. Phosphorylated claspin interacts with a phosphate-binding site in the kinase domain of Chk1 during ATR-mediated activation. *J. Biol. Chem.*, 278:46782–46788, 2003.

[91] E.N. Shiozaki, L. Gu, N. Yan, and Y. Shi. Structure of the BRCT repeats of BRCA1 bound to a BACH1 phosphopeptide: implications for signaling. *Mol. Cell*, 14:405–412, 2004.

[92] R.S. Williams, M.S. Lee, D.D. Hau, and J.N. Glover. Structural basis of phosphopeptide recognition by the BRCT domain of BRCA1. *Nat. Struct. Mol. Biol.*, 11:519–525, 2004.

[93] B.-B. S. Zhou and S. J. Elledge. The DNA damage response: putting checkpoints in perspective. *Nature*, 408:433–439, 2000.

[94] R. T. Abraham. Cell cycle checkpoint signaling through the ATM and ATR kinases. *Genes & Dev.*, 15:2177–2196, 2001.

[95] E.T. Boder and K.D. Wittrup. Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.*, 15:553–557, 1997.

[96] G.B. Dantzig. On the significance of solving linear programming problems with some integer variables. *Econometrica*, 28:30–44, 1960.

[97] K. E. Drexler. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. U.S.A.*, 78:5275–5278, 1981.

[98] C. O. Pabo. Molecular technology: Designing proteins and peptides. *Nature*, 301:200, 1983.

[99] J. Desmet, M. De Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542, 1992.

[100] R. F. Goldstein. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.*, 66:1335–1340, 1994.

[101] I. Lasters, M. De Maeyer, and J. Desmet. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. *Protein Eng.*, 8:815–822, 1995.

[102] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, 33:227–239, 1998.

[103] D. B. Gordon and S. L. Mayo. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comput. Chem.*, 19:1505–1514, 1998.

[104] D. B. Gordon and S. L. Mayo. Branch-and-terminate: A combinatorial optimization algorithm for protein design. *Structure*, 7:1089–1098, 1999.

[105] J. Mendes, A. M. Baptista, M. Arménia Carrondo, and C. M. Soares. Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model. *Proteins*, 37:530–543, 1999.

[106] M. De Maeyer, J. Desmet, and I. Lasters. The dead-end elimination theorem: Mathematical aspects, implementation, optimizations, evaluation, and performance. *Methods Mol. Biol.*, 143:265–304, 2000.

[107] L. L. Looger and H. W. Hellinga. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: Implications for protein design and structural genomics. *J. Mol. Biol.*, 307:429–445, 2001.

[108] P. Koehl and M. Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.*, 239:249–275, 1994.

[109] P. Koehl and M. Levitt. De novo protein design. I. In search of stability and specificity. *J. Mol. Biol.*, 293:1161–1181, 1999.

[110] P. Koehl and M. Levitt. Structure-based conformational preferences of amino acids. *Proc. Natl. Acad. Sci. U.S.A.*, 96:12524–12529, 1999.

[111] H. Kono and J. G. Saven. Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J. Mol. Biol.*, 306:607–628, 2001.

[112] C. Lee and S. Subbiah. Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.*, 217:373–388, 1991.

[113] H. W. Hellinga and F. M. Richards. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 91:5803–5807, 1994.

[114] P. S. Shenkin, H. Farid, and J. S. Fetrow. Prediction and evaluation of side-chain conformations for protein backbone structures. *Proteins*, 26:323–352, 1996.

[115] X. Jiang, E. J. Bishop, and R. S. Farid. A de novo designed protein with properties that characterize natural hyperthermophilic proteins. *J. Am. Chem. Soc.*, 119:838–839, 1997.

[116] X. Jiang, H. Farid, E. Pistor, and R. S. Farid. A new approach to the design of uniquely folded thermally stable proteins. *Protein Sci.*, 9:403–416, 2000.

[117] P. Tufféry, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rapid-determination of protein side-chain conformations. *J. Biomol. Struct. Dynam.*, 8:1267–1289, 1991.

[118] P. Tufféry, C. Etchebest, S. Hazout, and R. Lavery. A critical comparison of search algorithms applied to the optimization of protein side-chain conformations. *J. Comput. Chem.*, 14:790–798, 1993.

[119] P. Tufféry, C. Etchebest, and S. Hazout. Prediction of protein side chain conformations: A study on the influence of backbone accuracy on conformation stability in the rotamer space. *Protein Eng.*, 10:361–372, 1997.

[120] D. T. Jones. De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci.*, 3:567–574, 1994.

[121] J. R. Desjarlais and T. M. Handel. De novo design of the hydrophobic cores of proteins. *Protein Sci.*, 4:2006–2018, 1995.

[122] B. I. Dahiyat and S. L. Mayo. De novo protein design: Fully automated sequence selection. *Science*, 278:82–87, 1997.

[123] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim. High-resolution protein design with backbone freedom. *Science*, 282:1462–1467, 1998.

[124] B. Kuhlman, G. Dantas, G. C. Ireton, Varani G., B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302:1364–1368, 2003.

[125] J. R. Calhoun, H. Kono, S. Lahr, W. Wang, W. F. DeGrado, and J. G. Saven. Computational design and characterization of a monomeric helical dinuclear metalloprotein. *J Mol Biol*, 334:1101–1115, 2003.

[126] L.-P. Lee and B. Tidor. Barstar is electrostatically optimized for tight binding to barnase. *Nat. Struct. Biol.*, 8:73–76, 2001.

[127] Z. S. Hendsch and B. Tidor. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.*, 3:211–226, 1994.

[128] L.-P. Lee and B. Tidor. Optimization of electrostatic binding free energy. *J. Chem. Phys.*, 106:8681–8690, 1997.

[129] Z. S. Hendsch and B. Tidor. Electrostatic interactions in the GCN4 leucine zipper: Substantial contributions arise from intramolecular interactions enhanced on binding. *Protein Sci.*, 8:1381–1392, 1999.

[130] E. Kangas and B. Tidor. Optimizing electrostatic affinity in ligand–receptor binding: Theory, computation, and ligand properties. *J. Chem. Phys.*, 109:7522–7545, 1998.

[131] L.-P. Lee and B. Tidor. Optimization of binding electrostatics: Charge complementarity in the barnase–barstar protein complex. *Protein Sci.*, 10:362–377, 2001.

[132] N. C. J. Strynadka, H. Adachi, S. E. Jensen, K. Johns, A. Sielecki, C. Betzel, K. Sutoh, and M. N. G. James. Molecular-structure of the acyl-enzyme intermediate in beta-lactam hydrolysis at 1.7 angstroms. *Nature*, 359:700–705, 1992.

[133] N. C. J. Strynadka, S. E. Jensen, K. Johns, H. Blanchard, M. Page, A. Matagne, J-M Frère, and M. N. G. James. Structural and kinetic characterization of a β-lactamase-inhibitor protein. *Nature*, 368:657–660, 1994.

[134] N. C. J. Strynadka, S. E. Jensen, P. M. Alzari, and M. N. G. James. A potent new mode of inhibition revealed by the 1.7 Å X-ray crystallographic structure of the TEM-1–BLIP complex. *Nat. Struct. Biol.*, 3:290–297, 1996.

[135] P. J. Kraulis. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, 24:946–950, 1991.

[136] E. A. Merritt and D. J. Bacon. Raster3D: Photorealistic molecular graphics. *Meth. Enzymol.*, 277:505–524, 1997.

[137] L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein–protein recognition sites. *J. Mol. Biol.*, 285:2177–2198, 1999.

[138] A. T. Brünger and M. Karplus. Polar hydrogen positions in proteins: Empirical energy placement and neutron diffraction comparison. *Proteins*, 4:148–156, 1988.

[139] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.

[140] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.

[141] D. Sitkoff, K. A. Sharp, and B. Honig. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, 98:1978–1988, 1994.

[142] M. K. Gilson, K. A. Sharp, and B. H. Honig. Calculating the electrostatic potential of molecules in solution: Method and error assessment. *J. Comput. Chem.*, 9:327–335, 1988.

[143] K. A. Sharp and B. Honig. Electrostatic interactions in macromolecules: Theory and applications. *Annu. Rev. Biophys. Biophys. Chem.*, 19:301–332, 1990.

[144] K. A. Sharp and B. Honig. Calculating total electrostatic energies with the nonlinear Poisson–Boltzmann equation. *J. Phys. Chem.*, 94:7684–7692, 1990.

[145] J.J. Li, B.L. Williams, L.F. Haire, M. Goldberg, E. Walker, D. Durocher, M.B. Yaffe, S.P. Jackson, and S.J. Smerdon. Structural and functional versatility of the FHA domain in DNA-damage signaling by the tumor suppressor kinase Chk2. *Mol. Cell*, 9:1045–1054, 2002.

[146] M.F. Sanner, A.J. Olson, and J.C. Spehner. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers*, 38:305–320, 1996.

[147] M. Meyer, M. Desbrun, P. Schröder, and A.H. Barr. Discrete differential-geometry operators for triangulated 2-manifolds. In *VisMath '02*, 2002.

[148] S. Jones and J.M. Thornton. Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.*, 272:133–143, 1997.

[149] S. Jones, H.P. Shanahan, H.M. Berman, and J.M. Thornton. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.*, 31:7189–7198, 2003.

[150] O. Lichtarge, H. Yao, D.M. Kristensen, S. Madabushi, and I. Mihalek. Accurate and scalable identification of functional sites by evolutionary tracing. *J. Struct. Funct. Genomics*, 4:159–166, 2003.

[151] P. Chen, C. Luo, Y.L. Deng, K. Ryan, J. Register, S. Margosiak, A. Tempczyk-Russell, B. Nguyen, P. Myers, K. Lundgren, Kan C.C., and O'Connor P.M.

The 1.7 angstrom crystal structure of human cell cycle checkpoint kinase Chk1: Implications for Chk1 regulation. *Cell*, 100:681–692, 2000.

[152] W.S. Joo, P.D. Jeffrey, S.B. Cantor, M.S. Finnin, D.M. Livingston, and N.P. Pavletich. Structure of the 53BP1 BRCT region bound to p53 and its comparison to the Brca1 BRCT structure. *Genes Dev.*, 16:583–593, 2002.

[153] S. Orlicky, X.J. Tang, A. Willems, M. Tyers, and F. Sicheri. Structural basis for phosphodependent substrate selection and orientation by the SCFCdc4 ubiquitin ligase. *Cell*, 112:243–256, 2003.

[154] G. Waksman, S.E. Shoelson, N. Pant, D. Cowburn, and J. Kuriyan. Binding of a high-affinity phosphotyrosyl peptide to the Src Sh2 domain - Crystal-structures of the complexed and peptide-free forms. *Cell*, 72:779–790, 1993.

[155] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, 285:1735–1747, 1999.

[156] D.V. Bulavin, Y. Higashimoto, I.J. Popoff, W.A. Gaarde, V. Basrur, O. Porapova, E. Appella, and A.J. Fornace, Jr. Initiation of a G_2/M checkpoint after ultraviolet radiation requires p38 kinase. *Nature*, 411:102–107, 2001.

[157] S.-T. Kim, D.-S. Lim, C. E. Canman, and M. R. Kastan. Substrate specificities and identification of putative substrates of ATM family members. *J. Biol. Chem.*, 274:37538–37543, 1999.

[158] T. O'Neill, A. J. Dwyer, Y. Ziv, D. W. Chan, S. P. Lees-Miller, R. H. Abraham, J. H. Lai, D. Hill, Y. Shiloh, L. C. Cantley, and G. A. Rathbun. Utilization of oriented peptide libraries to identify substrate motifs selected by ATM. *J. Biol. Chem.*, 275:22719–22727, 2000.

[159] J. Bartek and J. Lukas. Chk1 and Chk2 kinases in checkpoint control and cancer. *Cancer Cell*, 3:421–429, 2003.

[160] D. Cortez, Y. Wang, J. Qin, and S.J. Elledge. Requirement of ATM-dependent phophorylation of BRCA1 in the DNA damage response to double-strand breaks. *Science*, 286:1162–1166, 1999.

[161] R.S. Tibbetts, D. Cortez, K.M. Brumbaugh, R. Scully, D. Livingston, S.J. Elledge, and R.T. Abraham. Functional interactions between BRCA1 and the checkpoint kinase ATR during genotoxic stress. *Genes Dev.*, 14:2989–3002, 2000.

[162] S. Banin, L. Moyal, S.Y. Shieh, Y. Taya, C.W. Anderson, L. Chessa, N.I. Smorodinsky, C. Prives, Y. Reiss, Y. Shiloh, and Y. Ziv. Enhanced phosphorylation of p53 by ATM in response to DNA damage. *Science*, 281:1674–1677, 1998.

[163] C.E. Canman, D.S. Lim, K.A. Cimprich, Y. Taya, K. Tamai, K. Sakaguchi, E. Appella, M.B. Kastan, and J.D. Siliciano. Activation of the ATM kinase by ionizing radiation and phosphorylation of p53. *Science*, 281:1677–1679, 1998.

[164] R.S. Tibbetts, K.M. Brumbaugh, J.M. Williams, J.N. Sarkaria, W.A. Cliby, S.Y. Shieh, Y. Taya, C. Prives, and R.T. Abraham. A role for ATR in the DNA damage-induced phosphorylation of p53. *Genes Dev.*, 13:152–157, 1999.

[165] N.D. Lakin, B.C. Hann, and S.P. Jackson. The ataxia-telangiectasia related protein ATR mediates DNA-dependent phosphorylation of p53. *Oncogene*, 18:3989–3995, 1999.

[166] C.A. Hall-Jackson, D.A.E. Cross, N. Morrice, and C. Smythe. ATR is a caffeine-sensitive, DNA-activated protein kinase with a substrate specificity distinct from DNA-PK. *Oncogene*, 18:6707–6713, 1999.

[167] B.D. Manning and L.C. Cantley. Hitting the target: Emerging technologies in the search for kinase substrates. *Sci. STKE*, 2002:pe49, 2002.

[168] L. L. Looger, M. A. Dwyer, J. J. Smith, and H. W. Hellinga. Computational

design of receptor and sensor proteins with novel functions. *Nature*, 423:185–190, 2003.

[169] J.M. Shifman and S.L. Mayo. Modulating calmodulin binding specificity through computational protein design. *J. Mol. Biol.*, 323:417–423, 2002.

[170] J.M. Shifman and S.L. Mayo. Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc. Natl. Acad. Sci. U.S.A.*, 100:13274–13279, 2003.

[171] T. Kortemme, L.A. Joachimiak, A.N. Bullock, A.D. Schuler, B.L. Stoddard, and D. Baker. Computational redesign of protein-protein interaction specificity. *Nat. Struct. Biol*, 11:371–379, 2004.

[172] Z. S. Hendsch, M. J. Nohaile, R. T. Sauer, and B. Tidor. Preferential heterodimer formation via undercompensated electrostatic interactions. *J. Am. Chem. Soc.*, 123:1264–1265, 2001.

[173] M.H. Ali, C.M. Taylor, G. Grigoryan, K.N. Allen, B. Imperiali, and A.E. Keating. Design of a heterospecific, tetrameric, 21-residue miniprotein with mixed alpha/beta structure. *Structure*, 13:225–234, 2005.

[174] D. N. Bolon, R. A. Grant, T. A. Baker, and R. T. Sauer. Specificity versus stability in computational protein design. *Proc. Natl. Acad. Sci. U.S.A.*, 102:12724–12729, 2005.

[175] L.A. Joachimiak, T. Kortemme, B.L. Stoddard, and D. Baker. Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. *J. Mol. Biol.*, 361:195–208, 2006.

[176] D. F. Green, A. T. Dennis, P.S. Fam, B. Tidor, and A. Jasanoff. Rational design of a new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide. *Biochem.*, in press.

[177] J. Ashworth, J.J. Havranek, C.M. Duarte, D. Sussman, R.J. Monnat, B.L. Stoddard, and D. Baker. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, 441:656–659, 2006.

[178] R. J. Hayes, J. Bentzien, M. L. Ary, Y. M. Hwang, J. M. Jacinto, J. Vielmetter, A. Kundu, and B. I. Dahiyat. Combining computational and experimental screening for rapid optimization of protein properties. *Proc. Natl. Acad. Sci. U.S.A.*, 99:15926–15931, 2002.

[179] J. G. Saven. Combinatorial protein design. *Curr. Opin. Struct. Biol.*, 12:453–258, 2002.

[180] M.C. Saraf, G.L. Moore, N.M. Goodey, V.Y. Cao, S.J. Benkovic, and C.D. Maranas. IPRO: An iterative computational protein library redesign and optimization procedure. *Biophys. J.*, 90:4167–4180, 2006.

[181] C.A. Bunton, D.R. Llewellyn, K.G. Oldham, and C.A. Vernon. 716. The reactions of organic phosphates. Part I. The hydrolysis of methyl dihydrogen phosphate. *J. Chem. Soc.*, pages 3574–3587, 1958.

[182] K. J. M. Hanf. *Protein Design with Hierarchical Treatment of Solvation and Electrostatics.* PhD thesis, Massachusetts Institute of Technology, 2002.

[183] J. A. Caravella. *Electrostatics and Packing in Biomolecules: Accounting for Conformational Change in Protein Folding and Binding.* PhD thesis, Massachusetts Institute of Technology, 2002.

[184] M. D. Altman. *Computational Ligand Design and Analysis in Protein Complexes Using Inverse Methods, Combinatorial Search, and Accurate Solvation Modeling.* PhD thesis, Massachusetts Institute of Technology, 2006.

[185] S.M. Lippow and B. Tidor. Unpublished data.

[186] R. L. Dunbrack, Jr. and M. Karplus. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.*, 230:543–574, 1993.

[187] R. L. Dunbrack, Jr. and F. E. Cohen. Bayesian statistical analysis of protein sidechain rotamer preferences. *Protein Sci.*, 6:1661–1681, 1997.

[188] W. E. Reiher III. *Theoretical studies of hydrogen bonding.* PhD thesis, Harvard University, Cambridge, MA, U.S.A., 1985.

[189] E. Neria, S. Fischer, and M. Karplus. Simulation of activation free energies in molecular systems. *J. Chem. Phys.*, 105:1902–1921, 1996.

[190] M. K. Gilson and B. Honig. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. *Proteins: Struct., Funct., Genet.*, 4:7–18, 1988.

[191] M. J. Nohaile, Z. S. Hendsch, B. Tidor, and R. T. Sauer. Altering dimerization specificity by changes in surface electrostatics. *Proc. Natl. Acad. Sci. U.S.A.*, 98:3109–3114, 2001.

[192] A. R. Fersht. The hydrogen bond in molecular recognition. *Trends Biochem. Sci.*, 12:301–304, 1987.

[193] A. R. Fersht, J.-P. Shi, J. Knill-Jones, D. M. Lowe, A. J. Wilkinson, D. M. Blow, P. Brick, P. Carter, M. M. Y. Waye, and G. Winter. Hydrogen bonding and biological specificity analysed by protein engineering. *Nature*, 314:235–238, 1985.

[194] Y.-J. Sun, J. Rose, B.-C. Wang, and C.-D. Hsiao. The structure of glutamine-binding protein complexed with glutamine at 1.94å resolution: comparisons with other amino acid binding proteins. *J. Mol. Biol.*, 278:219–229, 1998.

[195] M. Schaefer and M. Karplus. A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.*, 100:1578–1599, 1996.

[196] M. Schaefer, C. Bartels, and M. Karplus. Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model. *J. Mol. Biol.*, 284:835–848, 1998.

[197] M. Schaefer, C. Bartels, F. Leclerc, and M. Karplus. Effective atom volumes for implicit solvent models: Comparison between Voronoi volumes and minimum fluctuation volumes. *J. Comput. Chem.*, 22(15):1857–1879, 2001.

[198] E.T. Boder and K.D. Wittrup. Yeast surface display for directed evolution of protein expression, affinity, and stability. *Meth. Enzymol.*, 328:430–444, 2000.

[199] B.K. Kay, J. Winter, and J. McCaffery, editors. *Phage Display of Peptides and Proteins: A Laboratory Manual.* Academic Press, San Diego, 1996.

[200] M. A. Mena and P. S. Daugherty. Automated design of degenerate codon libraries. *Protein Eng. Des. Sel.*, 18:559–561, 2005.

[201] D. Bertsimas and J.N. Tsitsiklis, editors. *Introduction to Linear Optimization.* Athena Scientific, Belmont, MA, 1997.

[202] GAMS Development Corporation, Washington, DC. *GAMS: A User's Guide*, 2003.

[203] GAMS Development Corporation, Washington, DC. *GAMS: The Solver Manuals*, 2003.

[204] Cplex 8.0. In *GAMS: The Solver Manuals*, pages 65–96. GAMS Development Corporation, 2003.

[205] D.F. Green, B.A. Joughin, and B. Tidor. Design considerations for action-at-a-distance interactions that enhance binding affinity. *In preparation.*

[206] Y. Shaul and G. Schreiber. Exploring the charge space of protein–protein association: A proteomic study. *Proteins: Struct., Funct., Genet.*, 60:341–352, 2005.

[207] W.P. Russ, D.M. Lowery, P. Mishra, M.B. Yaffe, and R. Ranganathan. Natural-like function in artificial WW domains. *Nature*, 437:579–583, 2005.

[208] R.H. Schiestl and R.D. Gietz. High efficiency transformation of intact yeast cells using single stranded nucleic acids as a carrier. *Curr. Gen.*, 16:339–346, 1989.

[209] J.J. VanAntwerp and K.D. Wittrup. Thermodynamic characterization of affinity maturation: the D1.3 antibody and a higher-affinity mutant. *J. Mol. Recog.*, 11:10–13, 1998.

[210] M. Jäger, H. Nguyen, J.C. Crane, J.W. Kelly, and M. Gruebele. The folding mechanism of a β-sheet: the WW domain. *J. Mol. Biol.*, 311:373–393, 2001.

[211] J.P Tam. Synthetic peptide vaccine design: Synthesis and properties of a high-density multiple antigenic peptide system. *Proc. Natl. Acad. Sci. U.S.A.*, 85:5409–5413, 1988.